BAYESIAN SEMIPARAMETRICS

for modelling the

CLUSTERING OF EXTEME VALUES

MASTER'S THESIS

supervised by PROF. J. A. TAWN

under the responsibility of PROF. A. C. DAVISON

CHAIR OF STATISTICS

Thomas Lugrin — Spring 2013

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

STOR-i
excellence with impact

LANCASTER
UNIVERSITY

Contents

Introduction

The move from the study of maxima over a fixed period to the study of excesses over a high threshold allows for fitting models on more data. In the latter case however, we have to account for dependence between excesses, whereas yearly maxima can generally be assumed to be independent. The peaks over threshold approach suggests making inference only on maxima over clusters of exceedances, but there is no good method for specifying the clusters, and quantiles for long return periods can be significantly biased. Our goal is to focus on the subasymptotic model suggested by Eastoe and Tawn (2012) to develop new Bayesian semiparametric techniques to properly approach the problem, thus getting an estimation of the full uncertainty.

A simulation study shows the instability in estimated quantiles based on the peaks over threshold method compared to the subasymptotic model across different cluster definitions. An application on river peakflows is also presented and shows how the subasymptotic model, fitted with a novel semiparametric Bayesian method, can be applied to real data.

This work is divided into three main parts: an introductory part which provides an insight into multivariate extremes, with a particular attention to special kinds of dependence involved in this framework. In the second part we introduce a conditional multivariate model and develop a semiparametric Gibbs sampler to fit this particular model. The last part deals with the subasymptotic approach, meant to model short-range dependence of excesses over a threshold. We discuss further improvements and alternatives in the last section.

Introductory part

1 Background on multivariate extremes

Multivariate extremes were first studied in a bivariate context by Gumbel and Goldstein (1964) using two examples. The first dataset considered — oldest ages at death for both sexes — illustrated the case of general independence; the second — extremal floods at two gauging stations along the same river — involved overall dependence. The article extended the univariate block-maxima approach to inference on both margins, followed by an estimation of the overall dependence between them. The rise of the peaks over threshold method (Davison and Smith, 1990) led also to extensions to the bivariate case and inferences have been made on environmental as well as on financial data: for example, de Haan and de Ronde (1998) focused on estimating the probability for a dike to collapse due to extreme sea levels during wind storms, and Breymann *et al.* (2003) analysed the dependence structure on pairwise forex exchange rates. Our attention will be more focused on multivariate extremes, though most previous studies were limited to the bivariate setup.

Multivariate extreme values are not well-ordered. What then determines whether one d-dimensional observation is larger than another and should be considered as a maximum? As an example, we look at the bivariate case. It could be inappropriate to consider as maxima only pairs where both components dominate all other data pairs, i.e., $(X_{\max}, Y_{\max})$ is such that $X_{\max} > X_i, Y_{\max} > Y_i$ for all i, in case of extreme events rarely happening together. The structure underlying the common behaviour of both margins is of great interest. What is thus needed is a grasp of the upper tail joint behaviour, some knowledge about the overall structure of dependence and an understanding of the marginal — and maybe extremal — behaviour of each series.

The extreme value theorem is still valid in the multivariate framework. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n d-variate observations from the same underlying distribution F. Define $\boldsymbol{M}_n = \max\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ to be the componentwise maximum over this set, that is, $M_{n,i} := \max\{X_{i,1}, \ldots, X_{i,n}\}$, $i = 1, \ldots, d$. We have that $\Pr(\boldsymbol{M}_n < \boldsymbol{x}) = F^n(\boldsymbol{x})$. If there exist sequences $(\boldsymbol{a}_n) > \boldsymbol{0}$ and $(\boldsymbol{b}_n)$ such that

$$F^n(\boldsymbol{a}_n \boldsymbol{x} + \boldsymbol{b}_n) \to G(\boldsymbol{x}), \quad n \to \infty, \tag{1.1}$$

where the limit distribution G is non-degenerate in each margin, then G is a multivariate extreme value distribution. We say that F is in the *domain of attraction* of G. In the univariate context, G is known to be a member of one of three parametric families — depending on the nature of its tail. This does not hold in higher dimensions, however.

If the limit (1.1) holds, then, as convergence in distribution implies convergence in each margin, we have

$$F_i^n(a_{ni} x_i + b_{ni}) \to G_i(x_i), \quad i = 1, \ldots, d, \quad n \to \infty, \tag{1.2}$$

with F_i and G_i the margins of F and G respectively. This small result has deep consequences, for it justifies a parametric model to be used for inference on the margins.

The problem then divides into two parts: estimation of the margin parameters and inference on the dependence structure. This is supported by the fact that the distribution function F can be represented using a *copula* C, defined as a d-variate distribution with uniform margins:

$$\begin{aligned} C: \quad [0,1]^d &\longrightarrow [0,1] \\ \{F_1(x_1),\ldots,F_d(x_d)\} &\longmapsto F(\boldsymbol{x}). \end{aligned} \tag{1.3}$$

The copula representation is a way to remove any information related to the margins and it fully describes the dependence structure. It belongs to a broad class which cannot be covered by a parametric family. The main task is either to find a nonparametric inference procedure or a reasonable parametric subset of that nonparametric class: this subset should be large enough in order to be as unrestrictive as possible to avoid misspecification, but it should remain tractable and interpretable.

1.1 Inference on the margins

As an illustration for some examples and as an application of our discussion we use river flow data recorded in six different gauging stations in the United Kingdom. The last five recording stations are located on tributaries of the Thames (indirectly for the Ray and the Lambourn rivers which flow into the Cherwell and the Kennet rivers respectively, which in turn are tributaries of the Thames), the first being on the Thames itself. Their names and places, with some more details on their catchment geology, are listed in Table 1 (details provided by Marsh and Hannaford (2008)), and their geographical location is shown in Figure 1.

River	**Location**	**Catchment**	**BFI**
Thames	Eynsham	Oolitic Limestone and Oxford Clay	0.67
Ray	Grendon Underwood	Oxford Clay	0.18
Lambourn	Shaw	Chalk	0.96
Coln	Bibury	Oolitic Limestone	0.92
Mole	Gatwick Airport	Weald Clay	0.24
Ock	Abingdon	Chalk and Tertiary clays	0.63

Table 1 – The six gauging stations and their characteristics.

The data consist of daily mean flow in $m^3 s^{-1}$ measured for more than 4 decades up to the end of 2008. River flows are generally not directly recorded. A stage-discharge conversion is used to transform the measured river level or stage into a flow measurement (cf. the detailed description provided by Marsh and Hannaford (2008)). This is one of the errors introduced in the data, partly smoothed since daily means

Figure 1 – Location of the six gauging stations from which the data come: 1. the Thames at Eynsham, 2. the Ray at Grendon, 3. the Lambourn at Shaw, 4. the Coln at Bibury, 5. the Mole at Gatwick Airport, 6. the Ock at Abingdon.

are typically computed over 96 measurements (15-minute intervals originally). The river stage measures are also subject to errors, in particular during extreme events, as stated by Marsh and Hannaford (2008) in a comment on the weirs used for the Thames measurements at Eynsham, for which they mention "some bypassing at extreme discharges when [the] structure can be submerged". The structures — weirs, flumes — used for measurements are sometimes raised to face more extreme events, as for the Ray at Grendon Underwood after the 1964 flood. Perturbation factors are listed by the Institute of Hydrology (1980a). For the purpose of this study, we assume that the data are not biased due to these limitations.

A first comment when looking at the plotted data (two time series subsets are shown in Figure 2) is the remarkable difference in the shapes of the peaks. In one case the reaction to rainfalls is quick and sudden, while in the other case, an "echoing" phenomenon takes place, smoothing each peak on a much longer period. A direct effect in the latter case is that extremal events are harder to define.

These discrepancies in high flow events are explained by the response of the catchment to a rainfall. Chalky or oolitic ground stores the water like a sponge and releases it gradually. As an effect, it smooths and delays the additional flow due to precipitations. Conversely a clayey catchment tends to drain water more quickly to the river bed.

Another specificity of these data is the visible interpolations that have been already computed on some series, leading to surprising shapes in bivariate plots (an example is shown in Figure 3). This can be explained easily since we know that measurement issues are encountered especially during extreme events. These data have certainly been interpolated based on some extra knowledge. We therefore completed the small gaps corresponding to missing values by linear interpolation, assuming it

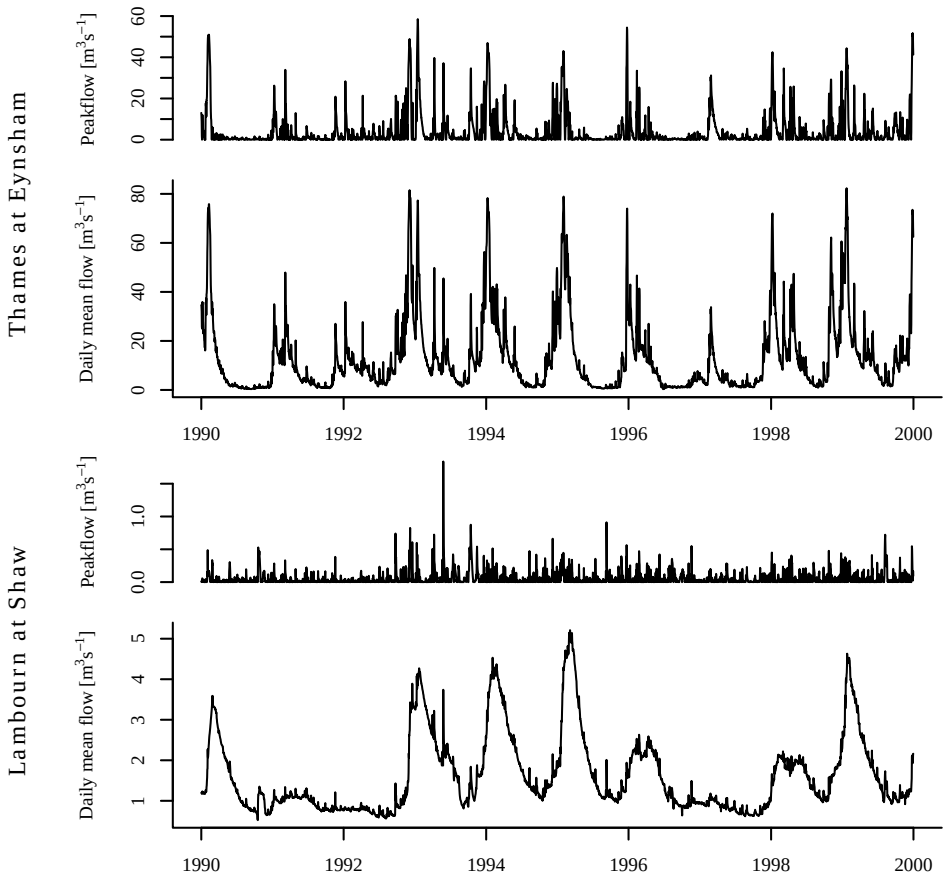

Figure 2 – Thames' and Lambourn's daily mean flows and daily peakflow in $m^3 s^{-1}$ at Eynsham and Shaw respectively, over one decade.

would not dramatically change the data structure. We will see below that, indeed, it does not affect inference on extremes.

As we have seen, the extreme peaks are differently spread through time, depending chiefly on the catchment geology. A statistic describing how much a river catchment is pervious has been developed by hydrological engineers (Institute of Hydrology, 1980b; Gustard *et al.*, 1992). It originates from research on determining periods of *low flow regime*. Further studies introduced the idea of separating flows into a *baseflow* component, which can be seen as the amount of water corresponding to low flow regime periods, and the *peakflow* – sometimes called *quickflow* –, which describes the unusually high flow levels. Figure 4 describes the process behind the flow separation: 5-day minima are first extracted from the daily flow series. We then consider triplets of minima $\{M_{i-1}, M_i, M_{i+1}\}$ and keep the central value M_i as a separation point only if $0.9M_i < \min\{M_{i-1}, M_{i+1}\}$. We compute a linear interpolation between the chosen points to define the separation line and, to ensure that this line is always at most as

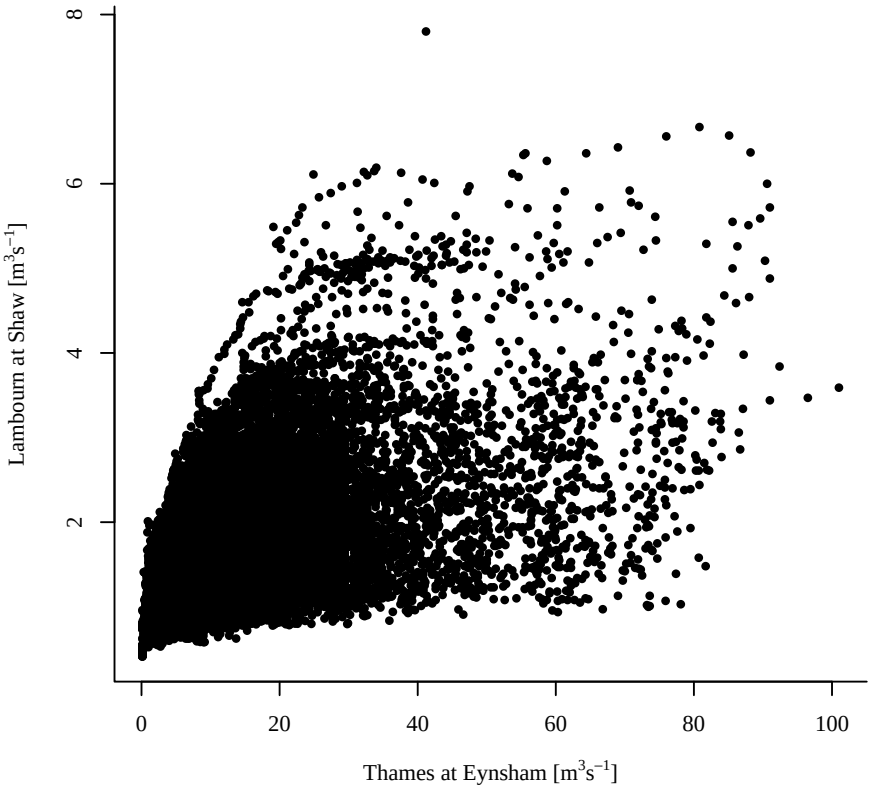

Figure 3 – Bivariate plot of data from the first (Thames) against the third flow time series (Lambourn). Missing values are ignored in this plot.

high as the entire flow, we lower baseflow values that are larger than the original flow measurements.

A summary value has been derived from this flow separation procedure. The *Base Flow Index* (BFI) is the ratio between the total baseflow and the total flow. It is a measure of how permeable the catchment is and thus how reactive the river flow is to a rainfall event. Obviously BFI $\in [0,1]$, but in practice the lowest values of BFI are 0.1, corresponding to a very flashy river, and the highest values are close to 1 for a very stable river (Gustard *et al.*, 1992). The link between this statistic, the catchment geological composition (Table 1) and the peak shapes (Figure 2) is undeniable. After removing the baseflow, we get much more similar peak shapes between the six time series (Figure 2 shows two of them). As we are interested in extreme events, we will focus throughout this work on peakflows only, and we leave the modelling of the whole flow, including the baseflow, for some further study.

In order to illustrate the previous discussion, we consider the peakflow data of the first and third gauging stations. Figure 5 shows the bivariate plot of these data. We notice that the linear structures completely disappeared from this plot, confirming the choice of replacing missing data in a simple way was not important.

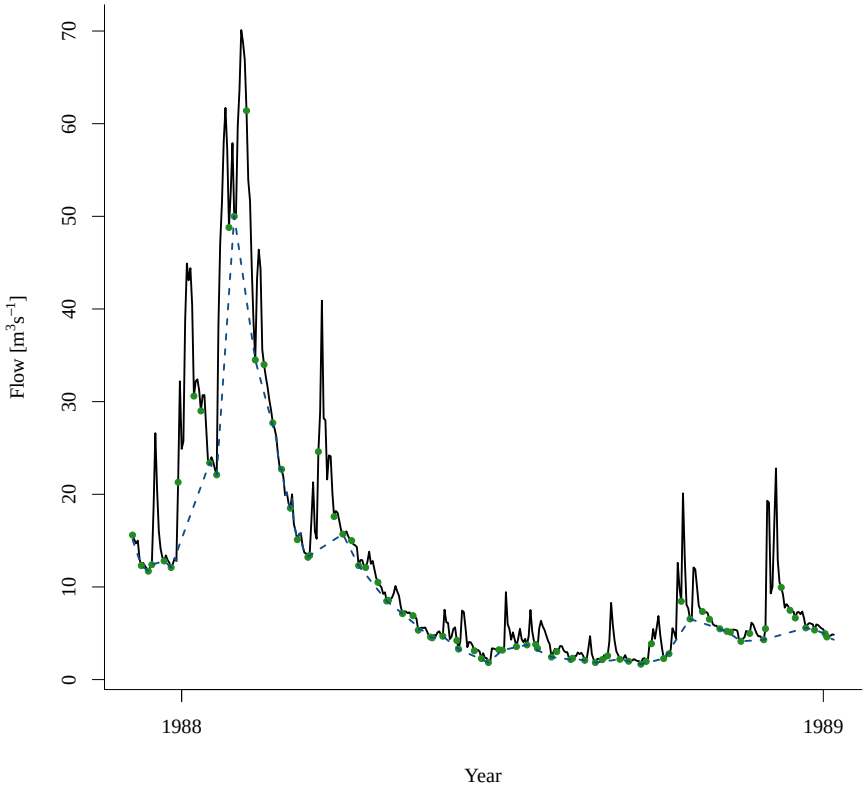

Figure 4 – Computation of the separation line between baseflow and peakflow. The green points indicate 5-day minima and the dashed blue line is the linearly interpolated separation line, after correction to be well-defined.

As seen before, the copula formulation (1.3) suggests separate inference on the margins and on the dependence structure. The marginals need not be uniform and can be more generally transformed into any continuous distribution. Here we standardise the margins to the Gumbel univariate distribution. The idea is to fix a high threshold u, beyond which the excesses are assumed to follow a generalised Pareto distribution (GPD). Below this threshold we adopt a nonparametric estimate for the distribution F_X of the observations $X_1,\ldots,X_n$. This leads to the semiparametric model of Coles and Tawn (1991, 1994):

$$\widehat{F}_X(x) := \begin{cases} \widetilde{F}_X(x), & x < u \\ 1-\{1-\widetilde{F}_X(x)\}\left(1+\xi\dfrac{x-u}{\sigma_u}\right)_+^{-1/\xi}, & x \geq u, \end{cases} \tag{1.4}$$

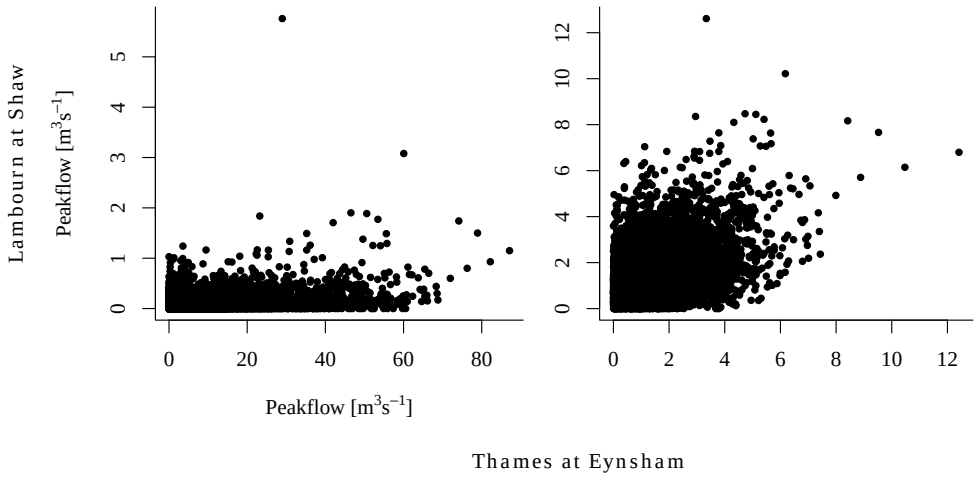

Figure 5 – Bivariate plots of peakflows on the original scale (left panel) and the Gumbel scale (right panel).

where $\widetilde{F}_X$ is the empirical distribution function of the data, ξ is the shape parameter, with appropriate limit interpretation when $\xi = 0$ and σ_u is the scale parameter, varying with u.

There are several complementary techniques used to choose a suitably high threshold in order to be as close as possible to the asymptotic regime, and low enough to ensure stability in the parameter estimates. In our example, we choose the 95% quantile as a threshold as the graphical diagnostics seem to indicate, corresponding to 23.8 and 0.24 m^3s^{-1} for the Thames and the Lambourn respectively. The inference is then made using optimisation methods (Nelder and Mead, 1965; Brent, 1973) available in R (R Development Core Team, 2012). That gives the results of Table 2. The shape parameter is in both cases significantly different from 0, and negative in the case of the Thames station, suggesting a bounded upper tail.

	$\widehat{\sigma}$	$\widehat{\xi}$
Thames	18.9 (0.008)	−0.43 (0.044)
Lambourn	0.14 (0.082)	0.23 (0.027)

Table 2 – Estimates of the scale (σ) and shape (ξ) parameters together with their standard errors for the generalised Pareto distribution fitted on the peakflow data of the Thames and the Lambourn.

1.2 Structure of dependence

In this section we assume that the margins $G_1,\ldots,G_d$ are known. Without loss of generality, we can thus assume them to be unit Fréchet by a simple application of the integral transform property, as shown below.

A general result provided by univariate extreme value theory is that the distribution G is itself in its own domain of attraction: if limit (1.1) holds, then for any $k \in \mathbb{N}^\star$ there exist $\boldsymbol{\alpha}_k > \mathbf{0}$ and $\boldsymbol{\beta}_k$ such that

$$G^k(\boldsymbol{\alpha}_k \boldsymbol{x} + \boldsymbol{\beta}_k) = G(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d.$$

Such a G distribution is called *max-stable*. The classes of multivariate extreme value and max-stable distributions actually coincide. In particular (Balkema and Resnick, 1977; Resnick, 1987), there exists a measure μ on $[-\infty,\infty)^d$ such that

$$G(\boldsymbol{x}) = \exp\left\{-\mu\left([0,\infty)^d \setminus [0,x]^d\right)\right\}, \quad \boldsymbol{x} \in [0,\infty]^d. \tag{1.5}$$

Several different *exponent measures* μ may satisfy equation (1.5), and Beirlant *et al.* (2004, chap. 8) give conditions on this measure to make it unique.

We write as $G_\star$ the multivariate distribution of the transformed vector

$$T_{\mathscr{F}}(\boldsymbol{X}) := \{-1/\log G_1(X_1),\ldots,-1/\log G_d(X_d)\}, \tag{1.6}$$

which is defined by

$$G_\star(\boldsymbol{x}) := G\left\{G_1^{\leftarrow}\left(\mathrm{e}^{-1/x_1}\right),\ldots,G_d^{\leftarrow}\left(\mathrm{e}^{-1/x_d}\right)\right\}, \quad \boldsymbol{x} \in (0,\infty)^d,$$

where $G_i^{\leftarrow}(x_i) := \inf\{x \in \mathbb{R} : G_i(x) > x_i\}$ is the generalised inverse of G_i. We can verify that $G_\star$ has unit Fréchet margins, since

$$\Pr\left\{T_{\mathscr{F}}(X_i) \le x\right\} = \Pr\left\{G_i(X_i) \le \mathrm{e}^{-1/x}\right\} = \mathrm{e}^{-1/x}, \quad 0 < x < \infty,\ i = 1,\ldots,d.$$

The distribution $G_\star$ has max-stable margins and is itself max-stable, and we even have that

$$G_\star^t(\boldsymbol{x}) = G_\star\left(t^{-1}\boldsymbol{x}\right), \quad \boldsymbol{x} \in \mathbb{R}^d,\ 0 < t < \infty. \tag{1.7}$$

With result (1.7), the exponent measure V corresponding to $G_\star$, defined as

$$V(\boldsymbol{x}) := -\log\left\{G_\star(\boldsymbol{x})\right\}, \quad x \in [0,\infty]^d,$$

is homogeneous of order -1, i.e., $V(t\boldsymbol{x}) = t^{-1}V(\boldsymbol{x})$, $0 < t < \infty$. A representation using pseudopolar coordinates is derived using this homogeneity property. These pseudopolar coordinates can be defined for arbitrary norms $\|\cdot\|_1$ and $\|\cdot\|_2$ in the corresponding mapping

$$\begin{aligned} \mathscr{T} : \mathbb{R}^{*,d} &\longrightarrow (0,\infty) \times \mathbb{S}_2^d \\ \boldsymbol{x} &\longmapsto (r,\boldsymbol{\omega}) = \left(\|\boldsymbol{x}\|_1, \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}\right), \end{aligned}$$

where $\mathbb{S}_2^d := \{\boldsymbol{\omega} \in \mathbb{R}^d : \|\boldsymbol{\omega}\|_2 = 1\}$ and $\mathbb{R}^*$ stands for $\mathbb{R} \setminus \{0\}$. As a change of norms is always possible through a simple formula (Beirlant *et al.*, 2004, chap. 8), we present without loss of generality the widely used sum-norm,

$$\|\boldsymbol{x}\|_{1,2} := \sum_{i=1}^{d} |x_i|.$$

The *spectral measure* $\widetilde{H}$ is defined on the $(d-1)$-dimensional unit simplex

$$S_d := \left\{\boldsymbol{\omega} \in [0,\infty)^d : \omega_1 + \cdots + \omega_d = 1\right\},$$

and the exponent measure is then given by

$$V(\boldsymbol{x}) = \int_{S_d} \max_{i=1,\ldots,d} \left\{\frac{\omega_i}{x_i}\right\} \mathrm{d}\widetilde{H}(\omega), \quad \boldsymbol{x} \in [0,\infty]^d. \tag{1.8}$$

We require the margins of $G_\star$ to be unit Fréchet, which is equivalent to

$$\int_{S_d} \omega_i \, \mathrm{d}\widetilde{H}(\boldsymbol{\omega}) = 1, \quad i = 1,\ldots,d, \tag{1.9}$$

and $\widetilde{H}$ must have mass d:

$$\int_{S_d} \mathrm{d}\widetilde{H}(\boldsymbol{\omega}) = d. \tag{1.10}$$

In the bivariate case, the standardised equations (1.8), (1.9) and (1.10) become

$$\begin{aligned} V(x_1,x_2) &= 2\int_0^1 \max\left\{\frac{\omega}{x_1}, \frac{1-\omega}{x_2}\right\} \mathrm{d}H(\omega), \\ \int_0^1 \omega \, \mathrm{d}H(\omega) &= \int_0^1 (1-\omega)\, \mathrm{d}H(\omega) = \frac{1}{2}, \qquad \int_0^1 \mathrm{d}H(\omega) = 1, \end{aligned} \tag{1.11}$$

and in this formulation $H = \widetilde{H}/d$ is the *spectral distribution*.

The need for partial derivatives of H has been addressed by Coles and Tawn (1991). These partial derivatives can be found through those of V:

$$\frac{\partial^m V}{\partial \boldsymbol{x}_I}(\boldsymbol{x}) = -\frac{1}{(\sum_{i\in I} x_i)^{m+1}} h_{m,I}\left(\frac{\boldsymbol{x}_I}{\sum_{i\in I} x_i}\right), \tag{1.12}$$

on $\{\boldsymbol{x} \in [0,\infty)^d : x_i = 0 \text{ if } i \notin I\}$, with $\boldsymbol{x}_I := \{x_i : i \in I\}$ and $I := \{i_1,\ldots,i_m\} \subset \{1,\ldots,d\}$. The densities $h_{m,I}$ are densities of edges, or subspaces of the simplex S_d. Specifically, they are defined on

$$S_{m,I} := \{\boldsymbol{\omega} \in S_d : \omega_i = 0 \text{ if } i \notin I\}.$$

The bivariate case gives a good example of result (1.12), with masses on the boundaries given by

$$H(\{0\}) = -\lim_{x_1 \to 0} \frac{\partial V}{\partial x_2}(x_1,x_2), \quad H(\{1\}) = -\lim_{x_2 \to 0} \frac{\partial V}{\partial x_1}(x_1,x_2),$$

and density within $(0,1)$ by

$$h(\omega) := h_{2,\{1,2\}} = -\frac{\partial^2 V}{\partial x_1 \partial x_2}(\omega, 1-\omega), \quad \omega \in (0,1).$$

Many parametric models for V have been proposed in the bivariate case, among which is the asymmetric mixed model (Tawn, 1988):

$$V(x_1,x_2) = \frac{x_1+x_2}{x_1x_2} - \frac{(\theta+\varphi)/x_1 + (2\varphi+\theta)/x_2}{x_1x_2(1/x_1+1/x_2)^2}, \tag{1.13}$$

with $\theta \geq 0$, $\theta+\varphi \leq 1$, $\theta+2\varphi \leq 1$ and $\theta+3\varphi \geq 0$. For this particular model, we get

$$H(\{0\}) = -\big(\theta+\varphi-1\big), \qquad H(\{1\}) = -\big(\theta+2\varphi-1\big),$$
$$h(\omega) = 2\big(\theta+3\varphi\omega\big), \quad \omega \in (0,1).$$

The independence case, corresponding to masses concentrated on $\{0,1\}$, arises when $\theta = \varphi = 0$. The asymmetry is directly translated into parameter φ, since it controls the slope of h.

Another way to summarise bivariate dependence is Pickands' function (Pickands, 1981):

$$A(t) := \int_0^1 \max\{\omega(1-t),(1-\omega)t\}\,\mathrm{d}H(\omega), \quad t \in [0,1], \tag{1.14}$$

or, by means of (1.11) and rearranging the terms,

$$A(t) = 1-t+2\int_0^t H([0,\omega])\,\mathrm{d}\omega, \quad t \in [0,1].$$

The exponent function V can be stated in terms of A as

$$V(x_1,x_2) = \frac{x_1+x_2}{x_1x_2} A\left(\frac{x_1}{x_1+x_2}\right).$$

The Pickands' function satisfies

(i) $\max(\omega,1-\omega) \leq A(\omega) \leq 1$ for $\omega \in [0,1]$,

(ii) $A(0) = A(1) = 1$,

(iii) $-1 \leq A'(0) \leq 0 \leq A'(1) \leq 1$,

(iv) $A''(\omega) \geq 0$, $\omega \in [0,1]$, at differentiable ω.

The bounds in (i) correspond to complete dependence and independence respectively. The spectral distribution H can in turn be expressed by means of A:

$$H([0,\omega]) = \begin{cases} \dfrac{1+A'(\omega)}{2}, & \omega \in [0,1), \\ 1, & \omega = 1, \end{cases}$$

with A' the right-hand derivative of A.

As we have to focus on the structure of dependence, let us have a closer look at the convergence of the copula (Beirlant *et al.*, 2004, chap. 8), as stated in (1.3). Write C_{F^n} as the copula of the sample maximum, so that we have

$$\begin{aligned} C_{F^n} &:= F^n\left\{\left(F_1^n\right)^{\leftarrow}(u_1),\ldots,\left(F_d^n\right)^{\leftarrow}(u_n)\right\} \\ &= F^n\left\{F_1^{\leftarrow}\left(u_1^{1/n}\right),\ldots,F_d^{\leftarrow}\left(u_d^{1/n}\right)\right\} =: C_F^n\left(u_1^{1/n},\ldots,u_d^{1/n}\right), \end{aligned}$$

since

$$\begin{aligned} \left(F_i^n\right)^{\leftarrow}(u_i) &:= \inf\left\{x \in \mathbb{R} : F_i^n(x) > u_i\right\} \\ &= \inf\left\{x \in \mathbb{R} : F_i(x) > u_i^{1/n}\right\} =: F_i^{\leftarrow}\left(u_i^{1/n}\right), \quad i = 1,\ldots,d. \end{aligned}$$

If F is in the domain of attraction of some multivariate extreme distribution G, then by continuity of C_G we get

$$\lim_{t\to\infty} C_F^t\left(u_1^{1/t},\ldots,u_d^{1/t}\right) = C_G(u_1,\ldots,u_d), \quad \boldsymbol{u} \in [0,1]^d,\ t \in \mathbb{R}. \tag{1.15}$$

Using the homogeneity of the exponent measure in terms of copulas, we get the approximation $C_F(\boldsymbol{u}) \approx C_G(\boldsymbol{u})$ for $\boldsymbol{u}$ sufficiently large in each of its components. We can translate that into an approximation for the distribution F by substituting $F_i(x_i)$ for u_i so that we can write, with a slight abuse of notation,

$$F(\boldsymbol{x}) \approx \exp\left[-V\left\{T_{\mathscr{F}}(x_1),\ldots,T_{\mathscr{F}}(x_d)\right\}\right], \tag{1.16}$$

for $\boldsymbol{x}$ close to its upper endpoint $\boldsymbol{x}^F := \sup\{\boldsymbol{x} : F(\boldsymbol{x}) < 1\}$. As guaranteed by the copula convergence in (1.15), the domain of attraction property may be rewritten with transformed marginals:

$$\lim_{t\to\infty} F^t\left\{T_{\mathscr{F}}^{\leftarrow}(t\boldsymbol{x})\right\} = G_\star(\boldsymbol{x}),$$

where $T_{\mathscr{F}}^{\leftarrow}(\boldsymbol{x}) := \left\{F_1^{\leftarrow}(\mathrm{e}^{-1/x_1}),\ldots,F_1^{\leftarrow}(\mathrm{e}^{-1/x_d})\right\}$. Applying $-\log(\cdot)$ to both sides and a first order Taylor expansion of the logarithm in the left-hand side term gives

$$\lim_{t\to\infty} t\left[1 - F\left\{T_{\mathscr{F}}^{\leftarrow}(t\boldsymbol{x})\right\}\right] = -\log G_\star(\boldsymbol{x}) = V(\boldsymbol{x}).$$

We then use copula C_F and replace t by $1/t$ to get

$$\lim_{t\downarrow 0}\frac{1}{t}\left\{1 - C_F\left(\mathrm{e}^{-t/x_1},\ldots,\mathrm{e}^{-t/x_d}\right)\right\} = \lim_{t\downarrow 0}\frac{1}{t}\Pr\left\{\bigcup_{i=1}^{d} F_i(X_i) > \mathrm{e}^{-t/x_i}\right\} = V(\boldsymbol{x}). \tag{1.17}$$

1.2.1 Nonparametric estimation

Results (1.16) and (1.17) are the basis for nonparametric estimation of the d-variate dependence structure. In particular, by setting $t = k/n$, $k \to \infty$ and $k/n \to 0$, when $n \to \infty$ equation (1.17) leads to an estimate for V:

$$\widehat{V}(\boldsymbol{x}) = \frac{1}{k}\sum_{i=1}^{n} \mathbf{1}\left\{\frac{k}{n} T_{\mathscr{F}}(\boldsymbol{X}_i) \not\leq \boldsymbol{x}\right\}. \tag{1.18}$$

In the bivariate case, this gives rise to a nonparametric estimate for Pickands' dependence function A. The derivation involves the structure variable

$$Z(\omega) := \max\left\{(1-\omega)T_{\mathscr{F}}(X_1), \omega T_{\mathscr{F}}(X_2)\right\}, \quad \omega \in [0,1],$$

whose estimates $\widehat{Z}(\omega)$ can be easily computed from a set of observations $\left(X_{i,1}, X_{i,2}\right)_{i=1}^{n}$. Using (1.18) and the fact that $A(\omega) = V\{1/(1-\omega), 1/\omega\}$ we obtain

$$\widehat{A}(\omega) = \frac{1}{k}\sum_{i=1}^{n} \mathbf{1}\left\{\widehat{Z}_i(\omega) > \frac{n}{k}\right\}, \quad \omega \in [0,1]. \tag{1.19}$$

However nothing ensures that this estimator is convex. Another way to find a nonparametric estimator for A is *via* the exponent distribution H. We transform data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ into pseudopolar coordinates and write

$$\widehat{R}_i := T_{\mathscr{F}}(X_{i,1}) + T_{\mathscr{F}}(X_{i,2}), \quad \widehat{W}_{i,j} := \frac{T_{\mathscr{F}}(X_{i,j})}{\widehat{R}_i}, \quad i = 1, \ldots, n,\ j = 1,2.$$

This leads to an estimator for H of the form

$$\widehat{H}(\cdot) = \frac{1}{k}\sum_{i=1}^{n} \mathbf{1}\left\{\widehat{R}_i > \widehat{R}_{(n-k)}, \widehat{W} \in \cdot\right\}, \tag{1.20}$$

where we choose $t = 1/\widehat{R}_{(n-k)}$, the $(k+1)$th largest $\widehat{R}_i$, instead of $t = k/n$, such that k observations remain above the threshold (Beirlant *et al.*, 2004, chap. 9). The estimator for A is derived from (1.14) by using (1.20):

$$\widehat{A}(t) = \frac{2}{k}\sum_{i=1}^{n} \mathbf{1}\left\{\widehat{R}_i > \widehat{R}_{(n-k)}\right\} \max\left\{(1-t)\widehat{W}_{i,1}, t\widehat{W}_{i,2}\right\}, \quad t \in [0,1]. \tag{1.21}$$

Estimators (1.19) and (1.21) can be modified in order for them to fulfil the constraints (i–iv), for example Beirlant *et al.* (2004) suggest

$$\widetilde{A}(t) = \max\left\{t, 1-t, \widehat{A}(t) + 1 - (1-t)\widehat{A}(0) - t\widehat{A}(1)\right\}, \quad t \in [0,1]. \tag{1.22}$$

Estimators (1.19) and (1.21) after having been modified through (1.22) are illustrated in Figure 7 for Thames and Lambourn data.

1.2.2 Parametric estimation

Since nonparametric estimators involve considering a region of the sample space with very few data and their upper endpoint is basically the largest observation, they may perform badly at asymptotic levels. This justifies the use of parametric estimation as a remedy for those drawbacks. We will focus on the *censored likelihood* method developed by Ledford and Tawn (1996) to derive estimators for A and V in the bivariate case.

We start from (1.16) with the following parametric form

$$F(\boldsymbol{x}) \approx \exp\left[-V\left\{\widehat{T}_{\mathscr{F}}(\boldsymbol{x}; \boldsymbol{\sigma}, \boldsymbol{\xi}); \boldsymbol{\theta}\right\}\right], \quad \boldsymbol{x} \geq \boldsymbol{u}, \tag{1.23}$$

where $\boldsymbol{u}$ is a suitably high threshold such that $1-F_j(u_j)$ is close to 0 for every $j = 1,\ldots,d$ and $\widehat{T}_{\mathscr{F}}(x)$ is defined as in (1.6) but using the marginal distribution model (1.4). Marginal and dependence parameters can be estimated jointly, enabling transfer of information between variables and allowing for constraints between marginal parameters, such as $\xi_i = \xi_j$, $i,j = 1,\ldots,d$. The corresponding likelihood for observations $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ is

$$L(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n;\sigma_1,\ldots,\sigma_d,\xi_1,\ldots,\xi_d,\boldsymbol{\theta}) = \prod_{i=1}^{n} L(\boldsymbol{X}_i). \tag{1.24}$$

Notice that model (1.23) takes only the region above $\boldsymbol{u}$ under consideration, that is, observations which are large in each of their coordinates. The idea of the censored likelihood method is to take into account the fact that some observations do not exceed the threshold, and this is achieved by censoring observations from below at u_j. In the bivariate case, the likelihood contributions in (1.24) depend on the region in which the corresponding observation falls in. We define four different regions

$$\begin{aligned} R_{00} &= \{(x_1,x_2) : x_1 \leq u_1, x_2 \leq u_2\}, \\ R_{01} &= \{(x_1,x_2) : x_1 \leq u_1, x_2 > u_2\}, \\ R_{10} &= \{(x_1,x_2) : x_1 > u_1, x_2 \leq u_2\}, \\ R_{11} &= \{(x_1,x_2) : x_1 > u_1, x_2 > u_2\}, \end{aligned} \tag{1.25}$$

and those are shown in Figure 6, together with the censoring principle, for which we forget about the actual value of the data coordinates under $\boldsymbol{u}$, as if it was shifted up to the threshold. The likelihood contributions in (1.24) are specific for each of these 4 regions, in the following way:

$$L(\boldsymbol{x}) \propto \begin{cases} F(u_1,u_2), & \boldsymbol{x} \in R_{00}, \\ \dfrac{\partial F}{\partial x_1}(x_1,u_2), & \boldsymbol{x} \in R_{10}, \\ \dfrac{\partial F}{\partial x_2}(u_1,x_2), & \boldsymbol{x} \in R_{01}, \\ \dfrac{\partial^2 F}{\partial x_1 \partial x_2}(\boldsymbol{x}), & \boldsymbol{x} \in R_{11}, \end{cases}$$

and depending on the specific parametric model chosen for V in (1.23), these contributions can be computed and the likelihood maximised.

In Figure 7 we present examples of estimated Pickands' function based on this censored likelihood, with the asymmetric mixed model (1.13) and the asymmetric logistic model (Tawn, 1988). The latter is expressed as

$$V(x_1,x_2) := \frac{1-\theta}{x_1} + \frac{1-\varphi}{x_2} + \left\{(\theta x_1)^{1/\alpha} + \left(\varphi x_2\right)^{1/\alpha}\right\}^{\alpha}, \quad x_1,\, x_2 > 0, \tag{1.26}$$

with $0 \leq \theta,\varphi,\alpha \leq 1$. Details about computations can be found in Appendix A.

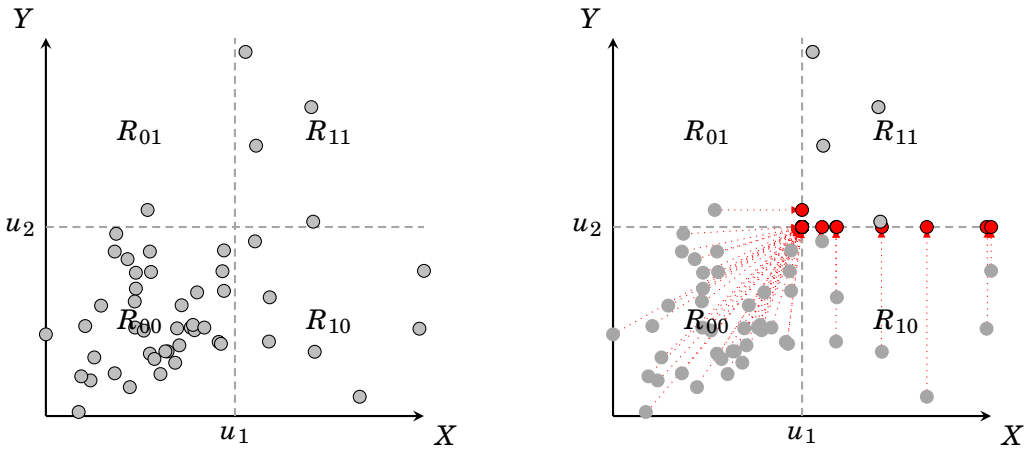

Figure 6 – Left panel: regions defined for the censored likelihood together with data on a Gumbel scale. Right panel: censoring principle, with red points referring to data that contribute only partially to the likelihood.

We now compare the four estimators described in the discussion above. The fact that estimator (1.19) is not always convex is apparent in Figure 7, whereas convexity is achieved by all other three estimators. The estimator provided under the pseudopolar coordinate setup seems to indicate more dependence than the other. The weakness of the naive estimator (1.19) resides in correction (1.22) which pulls the estimated Pickands' function upwards to make it equal 1 at $t = 0,1$, bringing it closer to independence than it indicates originally. No asymmetry is pointed out by the two parametric models, suggesting that their simpler symmetric version could be chosen instead.

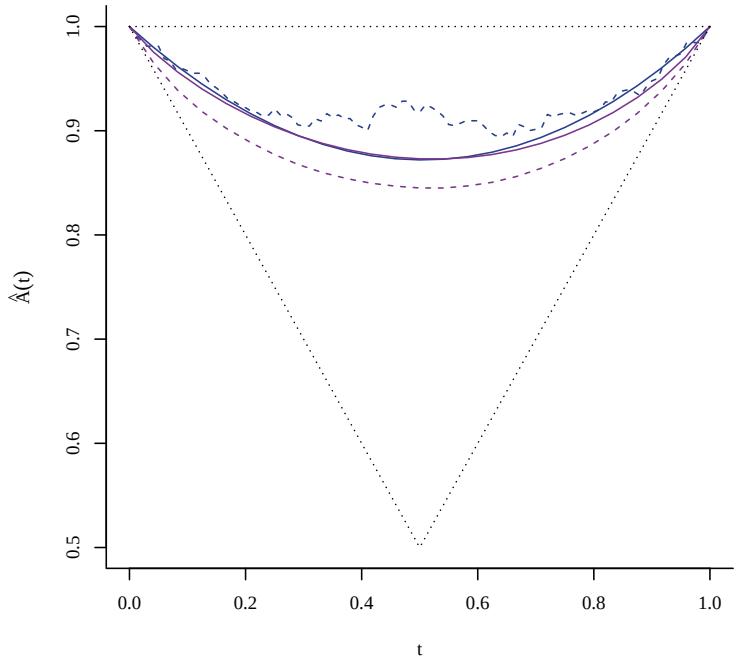

Figure 7 – Estimators of Pickands' dependence function on the Thames and Lambourn peak-flow data. Nonparametric estimators (dashed lines) modified in order to equal 1 at $t = 0,1$: blue one based on a naive estimator and violet one on pseudopolar coordinates. Parametric estimators (solid lines) using censored likelihood: blue one based on the asymmetric mixed model and violet one asymmetric logistic model. Extreme cases (dotted lines): independence — constant at 1 — and complete dependence — 'V' shape.

2 Dependence in extremes

Simple measures of correlation are not suitable when dealing with tails of distributions, since they reflect mostly their centre. There is a need for a tail dependence summary statistic, both to describe dependence within an extreme time series and to understand the joint behaviour of margins in a multivariate setup. In this section we present measures of extremal dependence, followed by a brief insight into how to estimate them with application on the peakflow data.

2.1 Measures of dependence

Our first look is towards dependence along a time series, for which we presented extremal convergence only in the case of independent and identically distributed random variables (cf. limit (1.2)). Leadbetter (1983) developed a condition under which weakly dependent distant extremes still converge in the same way as in (1.2). Consider a set of identically distributed variables $X_1,\dots,X_n$ and the *threshold sequence* u_n defined as verifying

$$n\{1-F(u_n)\} \to \tau, \quad n \to \infty,\ \tau > 0,$$

where $X_i \sim F$, $i = 1,\dots,n$. Let $I := \langle i_1 < \cdots < i_p \rangle$ and $J := \langle j_1 < \cdots < i_{p'} \rangle$ be non-overlapping increasing sequences of indices separated by a distance l, that is, $j_1 - i_p \geq l$. Condition $D(u_n)$ is then defined as

$$\left|F_{I,J}(u_n) - F_I(u_n)F_J(u_n)\right| < \alpha(n,l), \tag{2.1}$$

where $F_A(x)$ stands for $\Pr(X_i \leq x,\ i \in A)$ and $\alpha(n,l_n) \to 0$ as $n \to \infty$ with $l_n = o(n)$. Roughly stated, condition (2.1) ensures that for sufficiently distant extreme values, near-independence is verified.

This condition is hardly verifiable in practice, but seems to be a reasonable assumption in most cases. Raw flow data could be subject to long-range dependence, but $D(u_n)$ seems to be a suitable assumption when considering peakflow only. We now turn our interest to short-range dependence, for which Leadbetter (1983) derived another condition, albeit often unrealistic in practice.

When short-range dependence exists, we introduce the notion of *clusters* to describe large events occurring in limited periods of time: heavy rains lasting several consecutive days could for example be considered as one cluster. This is important since ignoring this clustering phenomenon would lead to overestimating occurrences of extreme events. A widely used measure of clustering is the *extremal index* that links the limiting distribution of $M_n := \max\{X_1,\dots,X_n\}$ and the distribution of the maximum built on independent replicates $\widetilde{M}_n := \{\widetilde{X}_1,\dots,\widetilde{X}_n\}$:

$$\Pr\left(\frac{M_n - b_n}{a_n} \leq x\right) \to G(x) \quad \Longleftrightarrow \quad \Pr\left(\frac{\widetilde{M}_n - b_n}{a_n} \leq x\right) \to \widetilde{G}(x), \quad n \to \infty, \tag{2.2}$$

with $(a_n) > 0$, (b_n) suitable normalising sequences, and the two limiting distributions are linked by the extremal index θ through $\left\{\widetilde{G}(x)\right\}^{\theta} = G(x)$.

Another definition of the extremal index leads to the interpretation of $1/\theta$ as being the limiting mean size of clusters. Following Leadbetter (1983) and writing Z_n as the number of exceedances of u_n in a block of length $r_n = o(n)$:

$$\theta^{-1} = \lim_{n\to\infty} \frac{r_n\{1-F(u_n)\}}{\Pr\left(M_{r_n} > u_n\right)} = \lim_{n\to\infty} \mathrm{E}(Z_n \mid Z_n \geq 1). \tag{2.3}$$

The last characterisation is based on considering exceedances close enough from each other as being in the same cluster:

$$\theta = \lim_{n\to\infty} \Pr\left(X_2 < u_n, \ldots, X_{r_n} < u_n \;\middle|\; X_1 > u_n\right) \tag{2.4}$$

After having looked at dependence within a time series, we consider the extremal dependence between two margins of a multivariate distribution. A natural measure of dependence at extreme levels for a pair (X_1, X_2) of F-distributed random variables is

$$\chi := \lim_{x\uparrow x^F} \Pr(X_1 > x \mid X_2 > x),$$

where, as before, x^F denotes the upper endpoint of distribution F. Asymptotic independence is reached when $\chi = 0$, whereas $0 < \chi \leq 1$ corresponds to asymptotic dependence. In the latter case, χ gives also an understanding of the degree of extremal dependence. A more convenient way of deriving χ is by using the copula formulation as follows:

$$\begin{aligned}
\lim_{x\uparrow x_F} \Pr(X_1 > x | X_2 > x) &= \lim_{F(x)\uparrow 1} \frac{\Pr\{F(X_1) > F(x),\, F(X_2) > F(x)\}}{\Pr\{F(X_2) > F(x)\}} \\
&= \lim_{u\uparrow 1} \frac{1-2u+C(u,u)}{1-u} \\
&= \lim_{u\uparrow 1} 2 - \frac{\log C(u,u)}{\log u} =: \lim_{u\uparrow 1} \chi(u),
\end{aligned} \tag{2.5}$$

allowing $\chi(u)$, $0 \leq u \leq 1$, to measure high quantiles dependence (Coles *et al.*, 1999). It turns out that $\chi(u)$ is constant at all levels for any distribution whose limit is a bivariate extreme value distribution. Non-constant values of $\chi(u)$ indicate that the extreme value class might be inappropriate to model the data.

The class of asymptotically independent distributions is poorly described by χ. Coles *et al.* (1999) introduced a complementary measure to provide more detailed information within this class:

$$\overline{\chi}(u) := \frac{2\log\Pr\{F(X_1) > u\}}{\log\Pr\{F(X_1) > u,\, F(X_2) > u\}} - 1 = \frac{2\log(1-u)}{\log \overline{C}(u,u)} - 1, \quad 0 \leq u \leq 1, \tag{2.6}$$

where $\overline{C}$ denotes the survivor copula and $-1 < \overline{\chi}(u) \leq 1$. The measure analogous to χ is defined as

$$\overline{\chi} := \lim_{u\uparrow 1} \overline{\chi}(u), \quad -1 \leq \overline{\chi} \leq 1.$$

Asymptotic dependence happens when $\overline{\chi} = 1$; in that case χ describes the strength of dependence within this class. Otherwise, $\overline{\chi} \in [-1,1)$ details the asymptotic independent case for which $\chi = 0$. The pair $(\chi,\overline{\chi})$ provides a complete summary of extremal dependence.

Up to this point, only bivariate dependence can be measured. Schlather and Tawn (2003) proposed a set of estimators for multivariate dependence known as the *extremal coefficients*. Let $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ be d-dimensional independent random variables with all marginal distributions being unit Fréchet. Under mild conditions, the distribution of the overall normalised maxima of variables indexed by $I \subseteq \{1,\ldots,d\}$ converges to a Fréchet distribution with parameter θ_I (not to be mistaken for the extremal index). Specifically,

$$\lim_{n\to\infty} \Pr\left(\max_{i\in I} \max_{j\in\{1,\ldots,n\}} \frac{X_{j,i}}{n} \leq x\right) = \lim_{n\to\infty} \left\{\Pr\left(\max_{j\in\{1,\ldots,n\}} \frac{X_{j,i}}{n} \leq x\right)\right\}^{\theta_I} = \exp\left(-\frac{\theta_I}{x}\right). \quad (2.7)$$

The whole set of relevant extremal coefficients has cardinality 2^d-1, which represents the number of choices for a non-empty subset in $\{1,\ldots,d\}$. We find a more helpful specification for θ_I by combining the integral form of V in relation (1.8) with definition (2.7):

$$\theta_I = \int_{S_d} \max_{i\in I}\{\omega_i\}\,\mathrm{d}H(\boldsymbol{\omega}),$$

with $1 \leq \theta_I \leq |I|$ and S_d the d-dimensional simplex.

2.2 Modelling dependence

As mentioned before, underestimation of short-range dependence can lead to overestimation of return levels, i.e., the size of a future extremal event on a given period of time. There is a necessity of *declustering* methods to cope with this sort of bias. A simple and effective approach was presented by Davison and Smith (1990). It is based on the fact that, under suitable mixing conditions, the asymptotic behaviour of cluster maxima is the same as for all exceedances of a threshold. The key point is in identifying clusters, in order to select their maximum value. In practice however the parameters of the generalised Pareto distribution estimated using this *peaks over threshold* method (POT) seem to present some bias.

Fawcett and Walshaw (2007) detailed a simpler inference using all exceedances of a high threshold that reduces this bias drastically. To account for underestimating uncertainty on the estimated parameters — due to a false independence working assumption — they complete this approach by inflating the confidence intervals by means of the covariance matrix of the likelihood gradient. This method estimates only the marginal distribution of the exceedances. Within the context of estimating the distribution of maxima, it corresponds to estimating $\widetilde{G}$ in equivalence (2.2) and thus does not help for specifying the maxima distribution of the dependent series. The first guess could be to estimate the extremal index separately (see below for suggestions of how to estimate it) and then modify the distribution accordingly. Within the context

of declustering, we will see later on how to greatly improve the understanding of the cluster maxima distribution using more sophistical tools (cf. §6 et seqq.).

Simple estimators are available for the extremal index θ, among which is the block estimator, derived from relation (2.3), that is the inverse of the mean number of exceedances per block, or

$$\widehat{\theta}_B = \frac{\sum_{i=1}^{\lfloor n/m \rfloor} \mathbf{1}\left\{\max\left(X_{(i-1)m+1},\ldots,X_{im}\right) > u\right\}}{\sum_{i=1}^{n} \mathbf{1}\{X_i > u\}}, \tag{2.8}$$

with m the block length. Relation (2.4) leads to the runs estimator

$$\widehat{\theta}_R = \frac{\sum_{i=1}^{n-m+1} \mathbf{1}\{\max(X_{i+1},\ldots,X_{i+m-1}) < u,\, X_i > u\}}{\sum_{i=1}^{n} \mathbf{1}\{X_i > u\}}. \tag{2.9}$$

Both (2.8) and (2.9) are easy to compute and generally consistent, but they heavily depend on the block length or runs length m, for the choice of which no general procedure exists (Ledford and Tawn (2003) give sufficient conditions that provide a value for m).

The above developments suggest the following inference for estimating the survivor distribution of extremes: estimate the shape parameter ξ and the scale parameter σ_u of the generalised Pareto distribution on observations above some suitable threshold u and get the conditional survivor distribution through

$$\Pr(X > x \mid X > u) = \left(1+\xi\frac{x-u}{\sigma_u}\right)_+^{-1/\xi}, \quad x > u, \tag{2.10}$$

and an estimate of θ with one of (2.8) or (2.9) can be used to get some insight into the cluster distribution.

A simple approach to estimating $\chi(u)$ and $\overline{\chi}(u)$, $u \in [0,1]$, is to transform the original observations $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ to uniform margins to obtain $\boldsymbol{U}_1,\ldots,\boldsymbol{U}_n$, to compute the empirical bivariate copula and survivor copula

$$\widehat{C}(\boldsymbol{u}) := \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}\{\boldsymbol{U}_i \leq \boldsymbol{u}\},$$
$$\widehat{\overline{C}}(\boldsymbol{u}) := \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}\{\boldsymbol{U}_i > \boldsymbol{u}\},$$

and to plug them in expressions (2.5) and (2.6) respectively. Figure 8 shows these estimates, fitted on the Thames and Lambourn peakflow data, together with 95% bootstrap confidence intervals (Fermanian *et al.*, 2004). The weird pattern for $u<0.25$ on both measures is explained by the large number of zero values within the series. It is not obvious how to draw conclusions on asymptotic dependence of our data based on these estimates, since neither does $\widehat{\chi}(u)$ seem to approach 0 nor does $\widehat{\overline{\chi}}(u)$ seem to have a limit at 1.

Coles *et al.* (1999) suggest parametric estimates of χ and $\overline{\chi}$ in order to complete the graphical diagnostics provided in Figure 8. These estimates are based on the characterisation of the joint survivor distribution by Ledford and Tawn (1996):

$$\Pr\{T_{\mathscr{F}}(X_1) > x,\ T_{\mathscr{F}}(X_2) > x\} \approx \mathscr{L}(x)x^{-1/\eta}, \quad \text{for large } x, \tag{2.11}$$

where $\mathscr{L}(x)$ is a slowly varying function as $x \to \infty$, i.e., $\lim_{x\to\infty} \mathscr{L}(tx)/\mathscr{L}(x) = 1$, $t > 0$, and $0 < \eta \leq 1$ is called the *coefficient of tail dependence*. The measure $\overline{\chi}$ is directly related to η since $\overline{\chi} = 2\eta - 1$. To enrich our nonparametric approach with the estimate of the limit measure $\overline{\chi}$, we use the Hill estimator (Hill, 1975) of η in the following way: define $Z_i = \min\{X_{i,1}, X_{i,2}\}$, $i = 1,\ldots,n$, so that $\Pr(Z > x)$ corresponds to the joint survivor distribution in (2.11). The estimator is then

$$\widehat{\eta}_k := \frac{1}{k}\sum_{i=1}^{k}\left(\log Z_{(n-k+i)} - \log Z_{(n-k)}\right), \tag{2.12}$$

where $Z_{(1)} \leq \cdots \leq Z_{(n)}$ are the order statistics of $Z_1,\ldots,Z_n$ and k has to be chosen small enough, typically such that $[k/n] = 5\%$.

To clarify the situation, we compute a parametric estimate of the tail of $\chi(u)$ based on the Ledford–Tawn model (2.11). We approximate the slowly varying function $\mathscr{L}$ at levels higher than $\widetilde{u}$ by a constant c, and we have the following model for $\chi(u)$:

$$\chi(u) := \Pr\left(T_{\mathscr{F}}(X_1) > u_{\mathscr{F}} \mid T_{\mathscr{F}}(X_2) > u_{\mathscr{F}}\right) = c u_{\mathscr{F}}^{1-1/\eta}, \quad u \in (\widetilde{u}, 1), \tag{2.13}$$

where $u_{\mathscr{F}} := -1/\log u$. We get an estimate for c through

$$\frac{\widehat{c}}{\widetilde{u}_{\mathscr{F}}^{1/\widehat{\eta}_k}} = \frac{k}{n},$$

since k/n is the empirical joint probability of exceeding threshold $\widetilde{u}$. From this we derive an estimator for $\chi(u)$ by introducing $\widehat{\eta}_k$ and $\widehat{c}$ into (2.13).

For the set of extremal coefficients θ_I, $I \subseteq \{1,\ldots,d\}$, Schlather and Tawn (2003) present a way of deriving self-consistent estimators. For simplicity however, we present only the 2-dimensional case with a naive estimator of $\theta_{\{1,2\}}$ which can be stated as $\widehat{\theta}_{\{1,2\}} = 2\widehat{A}(1/2)$, where $\widehat{A}$ is a suitable estimator of Pickands' function. Since $\chi = 2 - \theta_{\{1,2\}}$, we get further information about the asymptotic behaviour of $\chi(u)$, and this is also plotted in Figure 8. The confidence intervals for this last estimator could be much too optimistic and that the confidence bounds for the parametric estimator derived from the Ledford–Tawn model are much more reliable.

A conclusion based on this plot is that there is asymptotic independence with a positive association at extreme levels of the Thames' and Lambourn's peakflows, since χ is almost 0 and $\overline{\chi}$ is significantly less than 1.

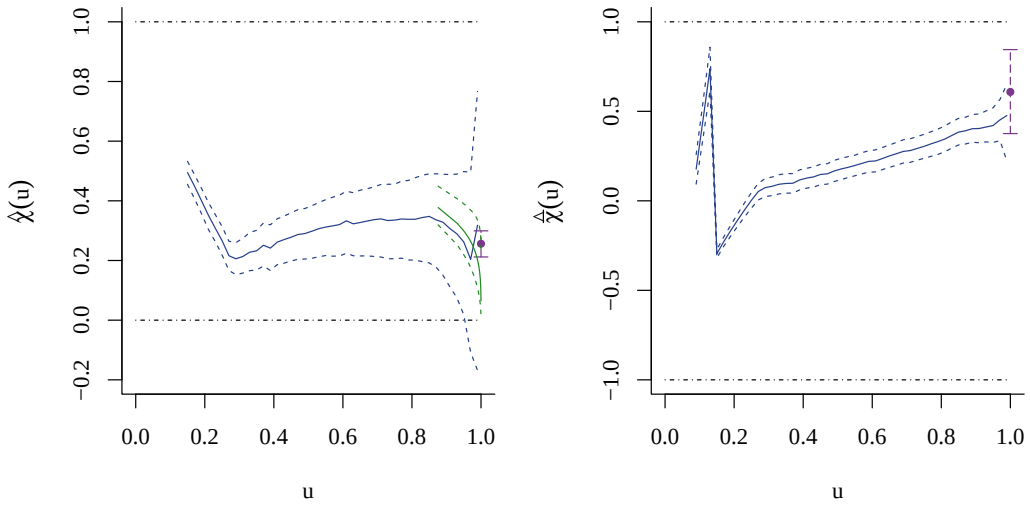

Figure 8 – Left panel: nonparametric estimation of (left panel) $\chi(u)$ as a blue solid line and of χ computed as $2-\theta_{\{1,2\}}$ in violet; the green solid line is a parametric estimate for $\chi(u)$ above a high threshold. Right panel: nonparametric estimation of $\overline{\chi}(u)$ and $\overline{\chi} = 2\eta - 1$ (same colour code). Dashed lines are 95% bootstrap pointwise confidence intervals and black broken lines represent the intervals in which χ and $\overline{\chi}$ can vary.

A model for multivariate extremes

3 Conditional multivariate extremes

In this section we work mostly on d-variate distributions F and G for which we need standard Gumbel margins. We define the transformations

$$T_{\mathscr{G}}(x) := -\log\{-\log J_i(x)\}, \quad \widehat{T}_{\mathscr{G}}(x) := -\log\{-\log \widehat{J}_i(x)\}, \quad i = 1,\dots,d, \tag{3.1}$$

where J_i stands for F_i or G_i, depending on the context, in the same way as — tacitly understood — for $T_{\mathscr{F}}$. Their hat-value is defined through (1.4). When x is multidimensional the transformation is meant componentwise with the suitable marginal distribution J_i for each component.

As may appear from previous sections, multivariate extreme inference is rarely extended beyond $d = 2$. The main issue arises from the need for parametric models to cover a large scope of extremal behaviours, from asymptotic dependence to asymptotic independence through all special cases lying in-between. The measures χ and $\overline{\chi}$ can give guidance for the model choice, but it has been outlined in a paper by Heffernan (2000) that most parametric models commonly used are very restrictive.

3.1 Heffernan–Tawn model

An innovative approach to inference for multivariate extremes was suggested by Heffernan and Tawn (2004). In their paper the authors describe a conditional approach that enables estimation of $\Pr(\boldsymbol{X} \in C)$ for any extreme set C in any dimension. Let $\boldsymbol{X}$ be a d-dimensional random variable with arbitrary marginal distributions $F_1,\dots,F_d$, such that there exist $(d-1)$-dimensional functions $\boldsymbol{a}_{|i}(x)$ and $\boldsymbol{b}_{|i}(x) > 0$ for which

$$\lim_{u\to\infty} \Pr\left[\frac{T_{\mathscr{G}}(\boldsymbol{X}_{-i}) - \boldsymbol{a}_{|i}\{T_{\mathscr{G}}(X_i)\}}{\boldsymbol{b}_{|i}\{T_{\mathscr{G}}(X_i)\}} \le \boldsymbol{z} \middle| X_i > u \right] = H_{|i}(\boldsymbol{z}), \tag{3.2}$$

where all marginal distributions of $H_{|i}$ are non-degenerate and $\boldsymbol{X}_{-i}$ denotes the vector $\boldsymbol{X}$ without its ith component. Thereafter we write the standardised $\boldsymbol{X}_{-i}$ as

$$\boldsymbol{Z}_{|i}(X_i) := \frac{T_{\mathscr{G}}(\boldsymbol{X}_{-i}) - \boldsymbol{a}_{|i}\{T_{\mathscr{G}}(X_i)\}}{\boldsymbol{b}_{|i}\{T_{\mathscr{G}}(X_i)\}}. \tag{3.3}$$

Under assumption (3.2) we get that $\sigma_u^{-1}(X_i - u)$, $\sigma_u{>}0$, is asymptotically conditionally independent of $\boldsymbol{Z}_{|i}$. This can be shown by considering a fixed $\widetilde{x} > 0$ with $x := u + \widetilde{x}\sigma_u$

and $u \to \infty$, as follows:

$$\begin{aligned}
\Pr\{\boldsymbol{Z}_{|i}(X_i) \leq \boldsymbol{z},\ \sigma_u^{-1}(X_i - u) > \widetilde{x} \mid X_i > u\} &= \\
&= \frac{\Pr\{\boldsymbol{Z}_{|i}(X_i) \leq \boldsymbol{z},\ X_i > x,\ X_i > u\}}{\Pr(X_i > u)} \\
&= \frac{\Pr\{\boldsymbol{Z}_{|i}(X_i) \leq \boldsymbol{z},\ X_i > x\}}{\Pr(X_i > u)} \\
&= \Pr\{\boldsymbol{Z}_{|i}(X_i) \leq \boldsymbol{z} \mid X_i > x\} \frac{\Pr(X_i > x)}{\Pr(X_i > u)} \\
&= \Pr\{\boldsymbol{Z}_{|i}(X_i) \leq \boldsymbol{z} \mid X_i > x\} \Pr\{\sigma_u^{-1}(X_i - u) > \widetilde{x} \mid X_i > u\} \\
&\to H_{|i}(\boldsymbol{z}) G(\widetilde{x}), \quad u \to \infty,
\end{aligned} \tag{3.4}$$

where G is the generalised Pareto distribution, i.e.,

$$G(x) = (1 + \xi x)_+^{-1/\xi}, \quad x > u,$$

with ξ the shape parameter as in (1.4).

Let $Z_{j|i}(x) := \left[T_{\mathscr{Q}}(X_j) - a_{j|i}\left\{T_{\mathscr{Q}}(x)\right\}\right] / b_{j|i}\left\{T_{\mathscr{Q}}(x)\right\}$, $j \neq i$, denote the jth component of $\boldsymbol{Z}_{|i}(x)$, with $a_{j|i}$ and $b_{j|i}$ the jth component of the vector functions $\boldsymbol{a}_{|i}$ and $\boldsymbol{b}_{|i}$ respectively. The distribution of each $Z_{j|i}(x)$, $j \neq i$, is written $H_{j|i}$ following the same logic as above, i.e., these are the marginal distributions of $H_{|i}$. We say that the elements of $\boldsymbol{X}_{-i}$ are mutually asymptotically conditionally independent given X_i if $H_{|i} = \prod_{j \neq i} H_{j|i}$.

The actual form of the functions $\boldsymbol{a}_{|i}$ and $\boldsymbol{b}_{|i}$ is simplified under the assumption that the elements of $\boldsymbol{X}$ are positively associated:

$$a_{j|i}(x) = \alpha_{j|i} x, \quad b_{j|i}(x) = x^{\beta_{j|i}}, \quad j \neq i,$$

with $0 \leq \boldsymbol{\alpha}_{|i} \leq 1$ and $-\infty < \boldsymbol{\beta}_{|i} \leq 1$. The Heffernan–Tawn conditional model can thus be stated as follows:

$$X_{j|i} = \alpha_{j|i} x + x^{\beta_{j|i}} Z_{j|i}(x), \quad X_i = x > u,\ j \neq i, \tag{3.5}$$

where u is the conditioning threshold, taken to be high enough in order to ensure that the conditional probability in (3.2) is close to its asymptotic regime. The conditional mean and variance are

$$\mu_{j|i}(x) = \alpha_{j|i} x + x^{\beta}_{j|i} \mu_{Z_{j|i}}, \quad \sigma^2_{j|i}(x) = x^{2\beta_{j|i}} \sigma^2_{Z_{j|i}}, \quad j \neq i, \tag{3.6}$$

giving the following translation of $\alpha_{j|i}$ and $\beta_{j|i}$ in terms of the dependence structure:

- $\alpha_{j|i} = 1$, $\beta_{j|i} = 0$ is the only case when (X_i, X_j) are asymptotically dependent, corresponding to $\chi > 0$ and $\overline{\chi} = 1$;
- if at least $0 < \alpha_{j|i} < 1$ or $\beta_{j|i} > 0$ holds, (X_i, X_j) fall in the case of asymptotic independence with positive extremal dependence, i.e., $\chi = 0$ and $0 < \overline{\chi} < 1$;

- if $\alpha_{j|i} = 0$ and $\beta_{j|i} \leq 0$ we get asymptotic independence with extremal near-independence, i.e., $\chi = \overline{\chi} = 0$.

The case of negative association is not covered here, but the function $\boldsymbol{a}_{|i}(x)$ can be extended in order to take this case into account. A modified approach introduced by Keef *et al.* (2013) involves a transformation to Laplace instead of Gumbel marginals; both tails are then exponentially distributed, so that all results for positive dependence also hold for negative dependence, with $\alpha_{j|i} < 0$. Figure 9 presents some examples of data generated from the Heffernan–Tawn bivariate model with values of $\alpha_{2|1}$ and $\beta_{2|1}$ corresponding to the cases listed above. Notice that in the case where $\beta_{2|1} < 0$, the conditional quantiles become closer as X_1 grows, which seems unlikely in most sets of data. We thus decide to set $\mathbf{0}$ as the lower bound for $\boldsymbol{\beta}_{|i}$.

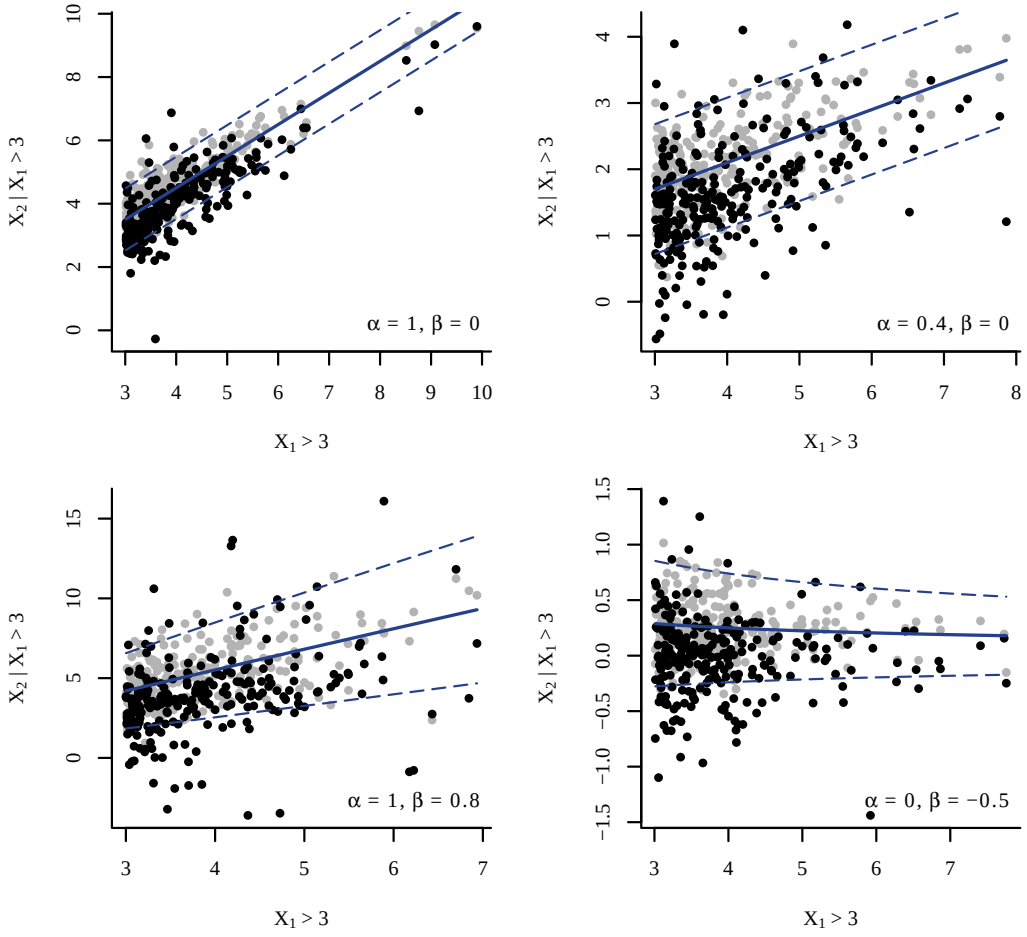

Figure 9 – Data generated from the Heffernan–Tawn model using $Z_{2|1} \sim \mathcal{N}(1/2, 1/4)$ — grey points — and $Z_{2|1} \sim$ Laplace(2) — black points. The solid blue line represents the conditional mean and the dashed ones the conditional 2.5% and 97.5% quantiles.

3.2 Two-step inference

Consider a sample of d-dimensional independent variables $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ from a standard Gumbel distribution. If it were not the case, estimating and applying $\widehat{T}_{\mathscr{G}}$ would bring the data back to that first assumption. The inference proposed by Heffernan and Tawn (2004) consists of a two-step method. Its advantage is that it is easy to implement. It is based on the working assumption that $Z_{j|i}$, $j \neq i$, are conditionally asymptotically independent and Gaussian with mean $\mu_{Z_{j|i}}$ and variance $\sigma^2_{Z_{j|i}}$. This assumption allows for a likelihood maximisation based on (3.6), for a high threshold u. The log-likelihood is of the form

$$\ell\left(\alpha_{j|i},\beta_{j|i},\mu_{j|i},\sigma^2_{j|i}\right) = -\sum_{k:X_{k,i}>u}\left[\log\sigma_{j|i}\left(X_{k,i}\right)+\frac{\left\{X_{k,j}-\mu_{j|i}\left(X_{k,i}\right)\right\}^2}{2\sigma_{j|i}\left(X_{k,i}\right)^2}\right], \quad j\neq i, \tag{3.7}$$

and by (3.6) we have a 4-parameter function. The first step is the maximisation of (3.7) from which we get estimates for $\boldsymbol{\alpha}_{|i}$, $\boldsymbol{\beta}_{|i}$, $\boldsymbol{\mu}_{Z_{|i}}$ and $\boldsymbol{\sigma}^2_{Z_{|i}}$. Since the assumed distribution of the $Z_{j|i}$'s was only convenient for the previous maximisation, we retain only the normalisation function estimators and build $H_{j|i}$, $j \neq i$ in a second step.

Given model (3.5) and the previous estimators, we can compute estimates for the residuals $\boldsymbol{Z}_{|i}$:

$$\widehat{Z}_{j|i}(X_{i,k}) = \frac{X_{j,k}-\widehat{\alpha}_{j|i}X_{i,k}}{X_{i,k}^{\widehat{\beta}_{j|i}}}, \quad k : X_{i,k} > u,\ j\neq i,$$

and from here build the empirical marginal distribution functions

$$\widehat{H}_{j|i}(z) = \frac{1}{n_u}\sum_{k:X_{i,k}>u}\mathbf{1}\left\{\widehat{Z}_{j|i}(X_{i,k})\leq z\right\}, \quad j\neq i,$$

with $n_u = \sum_{k=1}^{n}\mathbf{1}\left\{X_{i,k} > u\right\}$. The joint distribution $H_{|i}$ has an estimator of the form

$$\begin{aligned}\widehat{H}_{|i}(\boldsymbol{z}) &= \frac{1}{n_u}\sum_{k:X_{i,k}>u}\mathbf{1}\left\{Z_{j|i}\left(X_{i,k}\right) < z_j,\ j\neq i,\ j=1,\ldots,d\right\}\\ &= \frac{1}{n_u}\sum_{k:X_{i,k}>u}\prod_{\substack{j=1\\ j\neq i}}^{d}\mathbf{1}\left\{Z_{j|i}\left(X_{i,k}\right) < z_j\right\},\end{aligned}$$

which boils down to

$$\widehat{H}_{|i}(\boldsymbol{z}) = \prod_{\substack{j=1\\ j\neq i}}^{d}\widehat{H}_{j|i}(z_j),$$

under the assumption of marginal asymptotic independence.

This two-step estimation suffers however from underestimating the uncertainty of parameters estimated in the first step (de Carvalho and Ramos, 2012). Together with the Gaussian assumption about the residuals, we think that it could bring some misleading conclusions. This is why another inference method is proposed in what follows, but we shall first present the tools which it involves.

4 Semiparametric estimation *via* Gibbs sampling

In order to estimate the Heffernan–Tawn model in this new context, we first present the ingredients constituting the foundation of a nonparametric method. We then focus on how to adapt this method to our semiparametric problem.

4.1 Dirichlet process

Within the context of nonparametric estimation, the object we want to estimate is a distribution, thus the prior has to be a distribution over distributions. A widely used prior is the Dirichlet process, first presented by Ferguson (1973).

Let us first introduce some notation and the Dirichlet distribution, with which we can describe the Dirichlet process. Let the random vector $\boldsymbol{X}$ take values in a measurable space $(\mathscr{X},\mathscr{B})$. Let P be its unknown distribution, evolving in $(\mathscr{P},\mathscr{C})$, where $\mathscr{P}$ is the space of distributions on $(\mathscr{X},\mathscr{B})$ and $\mathscr{C}$ is the smallest σ-algebra generated by sets of the form $\{P \in \mathscr{P} : P(B) < v,\, B \in \mathscr{B},\, v \in [0,1]\}$.

We say that the random variable $\boldsymbol{X}$ is Dirichlet$(g_1,\ldots,g_p)$ if its density on the $(p-1)$-dimensional simplex S_p is

$$f(x_1,\ldots,x_{p-1}) = \frac{\Gamma\left(g_1+\cdots+g_p\right)}{\Gamma(g_1)\cdots\Gamma(g_p)}\left(1-\sum_{i=1}^{p-1}x_i\right)^{g_p-1}\prod_{i=1}^{p-1}x_i^{g_i-1}, \quad g_1,\ldots,g_p > 0. \tag{4.1}$$

Another more general definition states the Dirichlet distribution as follows: let $g_1,\ldots,g_p \geq 0$, such that $\sum_{i=1}^{p} g_i > 0$. Define Y_i as independent Gamma random variables with scale parameter 1 and shape parameter g_i, $i = 1,\ldots,p$. Writing $Y = \sum_{i=1}^{p} Y_i$, the distribution of $(X_1,\ldots,X_p)$, with $X_i := Y_i/Y$, $i = 1,\ldots,p$, is Dirichlet$(g_1,\ldots,g_p)$.

The Dirichlet process, as a prior for distributions, is a distribution on $(\mathscr{P},\mathscr{C})$ and can be defined through its realisations in $(\mathscr{P},\mathscr{C})$. Let P_0 be a distribution on $(\mathscr{X},\mathscr{B})$ — sometimes called the *baseline distribution*. Then the distribution of P is the Dirichlet process DP(P_0) if for every measurable partition $\{B_1,\ldots,B_p\}$ of $\mathscr{X}$ the distribution of the random vector $\{P(B_1),\ldots,P(B_p)\}$ is Dirichlet$\left\{P_0(B_1),\ldots,P_0(B_p)\right\}$.

A more general definition involves the *concentration parameter* $\gamma > 0$, with which the previous definition is simply changed such that for every measurable partition $\left\{B_1,\ldots,B_p\right\}$ of $\mathscr{X}$

$$\left\{P(B_1),\ldots,P(B_p)\right\} \sim \text{Dirichlet}\left\{\gamma P_0(B_1),\ldots,\gamma P_0(B_p)\right\}. \tag{4.2}$$

Given this formulation, we can derive the mean and variance of a DP(γP_0) realisation, i.e., of the random vector in (4.2):

$$\begin{aligned}
\mathrm{E}\{P(B_i)\} &= \frac{\gamma P_0(B_i)}{\sum_{j=1}^{p}\gamma P_0(B_j)} = P_0(B_i), \\
\mathrm{var}\{P(B_i)\} &= \frac{\gamma P_0(B_i)\left\{\sum_{j=1}^{p}\gamma P_0(B_j)-\gamma P_0(B_i)\right\}}{\left\{\sum_{j=1}^{p}\gamma P_0(B_j)\right\}^2\left\{\sum_{j=1}^{p}\gamma P_0(B_j)+1\right\}} = \frac{P_0(B_i)\{1-P_0(B_i)\}}{\gamma+1}.
\end{aligned}$$

The baseline distribution P_0 can thus be understood as the prior belief for P, and the concentration parameter γ reflects the prior uncertainty about P_0: the larger the parameter γ, the closer P will be to P_0.

4.2 Stick-breaking representation

The Dirichlet process can also be defined through a constructive approach, as introduced by Sethuraman (1994). As it is a central concept used in what will follow, we shall give a brief proof (deeply inspired by Sethuraman (1994)) of the fact that this definition is indeed equivalent to the one presented above.

Let $\mathscr{N}$ be the σ-algebra of all subsets of the integer set $\mathbb{N}$ and $\mathscr{E}$ the Borel σ-algebra on $[0,1]$. Let a set of random vectors $\{(V_k,\lambda_k,k) : k = 1,\ldots,K\}$ vary in $\{([0,1]\times\mathscr{X})^\infty\times\mathbb{N},(\mathscr{E}\times\mathscr{B})^\infty\times\mathscr{N}\}$, with the following properties:

$$\begin{aligned} V_1,V_2,\ldots &\overset{\text{iid}}{\sim} \text{Beta}(1,\gamma),\\ \lambda_1,\lambda_2,\ldots &\overset{\text{iid}}{\sim} P_0, \end{aligned}$$

with $(V_1,V_2,\ldots)$ independent of $(\lambda_1,\lambda_2,\ldots)$. Define the following probability weights:

$$\begin{aligned} w_1 &:= V_1,\\ w_k &:= V_k\prod_{i=1}^{k-1}(1-V_i), \quad k = 2,3,\ldots, \end{aligned}$$

and $\Pr(K = k \mid \boldsymbol{V},\boldsymbol{\lambda}) = w_k$. The *stick-breaking* representation of the Dirichlet process $\text{DP}(\gamma P_0)$ is then

$$P(B;\boldsymbol{V},\boldsymbol{\lambda}) := P(B) = \sum_{k=1}^{\infty} w_k\delta_{\lambda_k}(B), \quad B\in\mathscr{B}, \tag{4.3}$$

where $\delta_x(x) = 1$ and vanishes everywhere else.

We now give four preliminary results which are used to prove that the distribution of P is $DP(\gamma P_0)$.

For the first result, let $J = K-1$ and consider the random vector $(\boldsymbol{V}_{-1},\boldsymbol{\lambda}_{-1},J) := \{(V_2,V_3,\ldots),(\lambda_2,\lambda_3,\ldots),J\}$. We have the relation

$$\begin{aligned} P(B;\boldsymbol{V},\boldsymbol{\lambda}) &= w_1\delta_{\lambda_1}(B)+\sum_{k=2}^{\infty} w_k\delta_{\lambda_k}(B)\\ &= V_1\delta_{\lambda_1}(B)+(1-V_1)\sum_{k=2}^{\infty}\widetilde{w}_k\delta_{\lambda_k}(B)\\ &= V_1\delta_{\lambda_1}(B)+(1-V_1)P(B;\boldsymbol{V}_{-1},\boldsymbol{\lambda}_{-1}), \end{aligned} \tag{4.4}$$

where the weights in the right-hand side distribution are redefined as

$$\begin{aligned} \widetilde{w}_1 &:= \frac{w_2}{1-V_1} = V_2\\ \widetilde{w}_k &:= \frac{w_{k+1}}{1-V_1} = V_{k+1}\prod_{i=2}^{k}(1-V_i), \quad k = 2,3,\ldots, \end{aligned}$$

which shows that $(\boldsymbol{V}_{-1}, \boldsymbol{\lambda}_{-1})$ has the same distribution as $(\boldsymbol{V}, \boldsymbol{\lambda})$, and it is independent of (V_1, λ_1) by definition.

The second preliminary result, in (4.5), tells us about a property of Dirichlet distributions. Let $\boldsymbol{p}$ and $\boldsymbol{q}$ be non-negative, d-dimensional vectors, and $\boldsymbol{P}$ independent of $\boldsymbol{Q}$ random vectors having distributions Dirichlet($\boldsymbol{p}$) and Dirichlet($\boldsymbol{q}$) respectively. Let $p_{\text{sum}} := \sum_{j=1}^{d} p_j$ and $q_{\text{sum}} := \sum_{j=1}^{d} q_j$ and define the random variable V independent of $\boldsymbol{P}$ and $\boldsymbol{Q}$ and distributed as a Beta$(q_{\text{sum}}, p_{\text{sum}})$. Then

$$V\boldsymbol{Q} + (1-V)\boldsymbol{P} \;\sim\; \text{Dirichlet}(\boldsymbol{p}+\boldsymbol{q}). \tag{4.5}$$

We use the same notation for the following third preliminary result. Let $\overline{p}_j := p_j/p_{\text{sum}}$, $j = 1,\ldots,d$, and denote by $\boldsymbol{e}_j$ the canonical vector with zeroes everywhere except for its jth element being 1. Then

$$\sum_{j=1}^{d} \overline{p}_j \,\text{Dirichlet}(\boldsymbol{p}+\boldsymbol{e}_j) \;=\; \text{Dirichlet}(\boldsymbol{p}). \tag{4.6}$$

The last preliminary result needed is about the uniqueness of the solution for a specific type of distributional equations. If V, P and Q are random variables with P independent of the other ones, and $\Pr(V = 1) < 1$, then there is a unique P solution to the equation

$$P \;=\; Q + VP. \tag{4.7}$$

We now show that for any partition $\{B_1,\ldots,B_p\}$, $\boldsymbol{P} := \{P(B_1),\ldots,P(B_p)\}$, with elements defined as in (4.3), is Dirichlet$\{\gamma P_0(B_1),\ldots,\gamma P_0(B_p)\}$.

Let $\boldsymbol{D} := \{\delta_{\lambda_1}(B_1),\ldots,\delta_{\lambda_1}(B_p)\}$. Then

$$\Pr(\boldsymbol{D} = \boldsymbol{e}_j) = \Pr(\delta_{\lambda_1}(B_j) = 1,\, \delta_{\lambda_1}(B_i) = 0,\, i \neq j) = \Pr(\lambda_1 \in B_j) = P_0(B_j), \tag{4.8}$$

since $\{B_1,\ldots,B_p\}$ is a partition. From (4.4) we have that

$$\boldsymbol{P} \;=\; V_1\boldsymbol{D} + (1-V_1)\boldsymbol{P}, \tag{4.9}$$

where $V_1 \sim \text{Beta}(1,\gamma)$ as before. We now verify that $\boldsymbol{P}$ can be replaced by the distribution Dirichlet$\{\gamma P_0(B_1),\ldots,\gamma P_0(B_p)\}$ in the distributional equation (4.9). We compute the right-hand side density in (4.9) by first conditioning on $\boldsymbol{D}=\boldsymbol{e}_j$ and then integrating out on $\boldsymbol{D}$. We have

$$\begin{aligned}\left(V_1\boldsymbol{D} + (1-V_1)\boldsymbol{P} \;\middle|\; \boldsymbol{D} = \boldsymbol{e}_j\right) = \\ V_1 \,\text{Dirichlet}(\boldsymbol{e}_j) + (1-V_1)\,\text{Dirichlet}\{\gamma P_0(B_1),\ldots,\gamma P_0(B_p)\},\end{aligned} \tag{4.10}$$

since the distribution Dirichlet($\boldsymbol{e}_j$) is degenerate in $\boldsymbol{e}_j$. By (4.5), the distribution in (4.10) is Dirichlet$\left[\{\gamma P_0(B_1),\ldots,\gamma P_0(B_p)\} + \boldsymbol{e}_j\right]$. We now integrate over the distribution of $\boldsymbol{D}$, and thanks to (4.8), this distribution is equivalent to

$$\sum_{j=1}^{p} P_0(B_j)\,\text{Dirichlet}\left[\{\gamma P_0(B_1),\ldots,\gamma P_0(B_p)\} + \boldsymbol{e}_j\right],$$

which is Dirichlet$\{\gamma P_0(B_1),\ldots,\gamma P_0(B_p)\}$ by the result stated in (4.6). It remains to show that the solution to (4.9) is unique. We have that $\boldsymbol{P}$ is independent of $(V_1,\boldsymbol{D})$, as mentioned after (4.4), and $\Pr(V_1 = 0) < 1$ if γ is strictly positive. As this is an initial assumption, the uniqueness of the solution to the distributional equation (4.7) ensures that (4.9) admits only one solution, that is, P has distribution DP(γP_0).

Before establishing a model for this particular prior distribution, we explain the intuitive understanding of a stick-breaking realisation. Given a stick of length 1, choose a location where to break it, according to a Beta$(1,\gamma)$. This is the first weight w_1. Take the remaining stick, elongate it such that it has length 1, and break it again, randomly. The cut part, shrunk back to its original size — before elongation —, gives w_2. Elongate the remaining part, and repeat the previous steps to get the following weights. Figure 10 depicts the stick-breaking process in a more intelligible way.

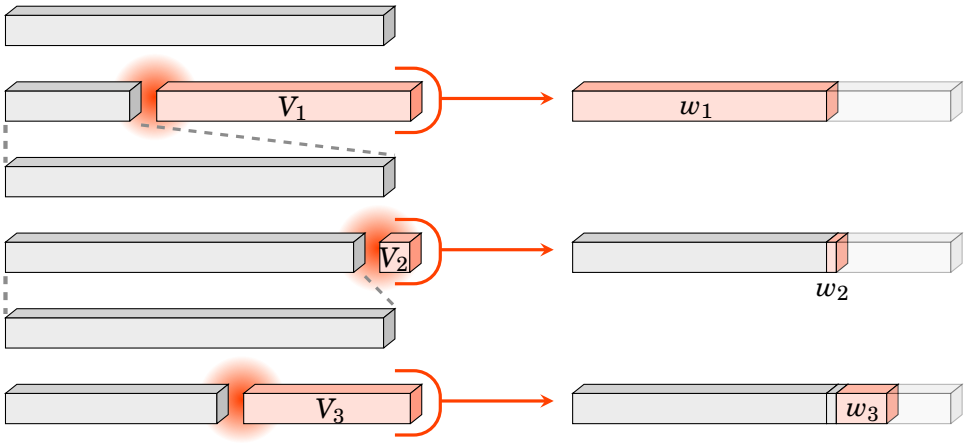

Figure 10 – Illustration of the stick-breaking procedure. Left panel: stick breaks, each time zooming on the remaining part. Right panel: weights computed from the breaks performed in the left panel.

4.3 Dirichlet mixture models

A simple model structure involving a Dirichlet process prior arises when estimating a mixture of distributions. The purpose could be the determination of clusters, or components, in a given dataset or a generalisation of kernel smoothing densities. The latter case will be detailed later on.

If $\{X_1,\ldots,X_n\}$ represents a set of observations, the model has the form

$$\begin{aligned} X_i \mid \lambda_i &\overset{\text{ind}}{\sim} H(\cdot;\lambda_i), \quad i = 1,\ldots,n, \\ \lambda_i \mid P &\overset{\text{iid}}{\sim} P, \quad i = 1,\ldots,n, \\ P &\sim \text{DP}(\gamma P_0), \end{aligned} \tag{4.11}$$

with the same notation as before.

If we are interested in the structure of mixture components within $\{X_1,\dots,X_n\}$, then we may introduce an index variable c_i, $i = 1,\dots,n$, which maps each observation to its respective component. We update model (4.11) as follows:

$$\begin{aligned} X_i \mid c_i, \boldsymbol{\lambda} &\overset{\text{ind}}{\sim} H(\cdot;\lambda_{c_i}), \quad i = 1,\dots,n, \\ \lambda_k \mid P &\overset{\text{iid}}{\sim} P, \quad k = 1,2,\dots, \\ P &\sim \text{DP}(\gamma P_0), \end{aligned} \tag{4.12}$$

and the indices c_i, $i = 1,\dots,n$, give the information about which observation belongs to which component.

In the case where we want to generalise the kernel smoothing procedure, we could take H to be a normal distribution in (4.12) and after fixing a variance σ_H^2 for this distribution, the λ_i's would represent the mean of each component. But this still involves estimating the variance σ_H^2, which brings us back to the problem of estimating a bandwidth. We can thus make σ_H^2 vary across the components, leading to 2-dimensional $\boldsymbol{\lambda}_k$'s, specifically $\boldsymbol{\lambda}_i = (\mu_k, \sigma_k^2)$, $k = 1,2,\dots$, with μ_{c_i} and $\sigma_{c_i}^2$ the mean and variance of component c_i, $i = 1,\dots,n$. An application of such a generalised kernel density estimation is presented in the next section.

4.4 Gibbs sampler

As we can observe from the definition of the stick-breaking representation (4.3), there is an unbounded amount of weights w_i to compute in order to get a realisation of distribution P in (4.3). Two approaches have been exploited so far: the first, introduced by Escobar (1994), exploits the Dirichlet process representation of Blackwell and MacQueen (1973) and leads to the marginal representation

$$\begin{aligned} \lambda_1 &\sim P_0, \\ \lambda_{k+1} \mid \lambda_k,\dots,\lambda_1 &\sim \frac{\gamma}{\gamma+k} P_0 + \frac{1}{\gamma+k} \sum_{j=1}^{k} \delta_{\lambda_j}. \end{aligned}$$

MacEachern and Müller (1998) and Neal (2000) gave what are considered as state-of-the-art algorithms based on this representation.

Ishwaran and Zarepour (2000) proposed a second approach, with the aim of sampling directly from the posterior distribution. Their method is based on the truncation of the stick-breaking representation (4.3), which allows for a complete reformulation of model (4.12) in terms of random variables. Let P be rewritten as

$$P = \sum_{k=1}^{N} w_k \delta_{\boldsymbol{\lambda}_k},$$

with N a fixed non-zero integer which determines the upper bound of the number of mixture components. From now on, the $\boldsymbol{\lambda}_k$'s are considered more generally as vectors

of parameters. To ensure that the weights still sum to one, we add the constraint $V_N = 1$. As can be seen with a simple recursion argument,

$$\sum_{k=1}^{N} w_k := V_1 + \sum_{k=2}^{N} V_k \prod_{j=2}^{k-1}(1-V_j) = 1 - \prod_{k=1}^{N}(1-V_k),$$

and by imposing $V_N = 1$ the last product on the right-hand side vanishes. The truncation assumption is supported by a convergence result on the posterior distribution of $\boldsymbol{c}$ (Ishwaran and James, 2002); the truncated posterior for $\boldsymbol{c}$ converges to its non-truncated version as $O\left[n\exp\left\{-(N-1)/\gamma\right\}\right]$. It implies that for values of γ up to 3 and $1,000$ observations, we have an accurate approximation — error of order 10^{-4} — with $N = 50$.

The finite approximation of the stick-breaking representation is used in Figure 11 to represent the weights w_k, $k = 1,\ldots,N$, depending on the value of the concentration parameter γ. Higher values for γ produce a distribution P closer to its baseline P_0. The plots can be thought of as the posterior densities for the mean and variance in a generalised kernel density estimation, for which an example will be presented afterwards.

To illustrate the discussion above, we consider a dataset of galaxy speeds, recorded in the region of Corona Borealis (Postman *et al.*, 1986), and restrict ourself to the dataset studied by Roeder (1990), composed of 82 galaxies from 6 well-separated conic sections of space. This dataset has been reanalysed in several works dealing with Dirichlet processes (Ishwaran and James, 2002; Fearnhead, 2004).

The model we consider is based on the work by Ishwaran and James (2001), and can be developed from (4.12) in the following way:

$$\begin{aligned} X_i \mid c_i, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 &\overset{\text{ind}}{\sim} \mathcal{N}\left(\mu_{c_i}, \sigma^2_{c_i}\right), \quad i = 1,\ldots,n, \\ c_i \mid \boldsymbol{w} &\overset{\text{iid}}{\sim} \sum_{k=1}^{N} w_k \delta_k, \quad i = 1,\ldots,n, \\ \mu_k &\overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2_{\mu}\right), \quad k = 1,\ldots,N, \\ \sigma^2_k &\overset{\text{iid}}{\sim} \text{Inv-Gamma}(\nu_1, \nu_2), \quad k = 1,\ldots,N, \end{aligned} \tag{4.13}$$

where σ^2_{μ} is chosen sufficiently large to get an uninformative prior, and (ν_1, ν_2) is typically $(2,2)$ to make the model prefer more components with smaller variances rather than few, non-informative, and wide components.

The truncation of the stick-breaking representation is the key point to allow for computing the posterior distributions involved in model (4.13). As proposed by Ishwaran and James (2002), we can increase the flexibility of this model by introducing a further level in the hierarchy. Specifically, we add hyperpriors on the concentration parameter γ and on the means $\boldsymbol{\mu}$ to cope with non-centred data. All together, the first

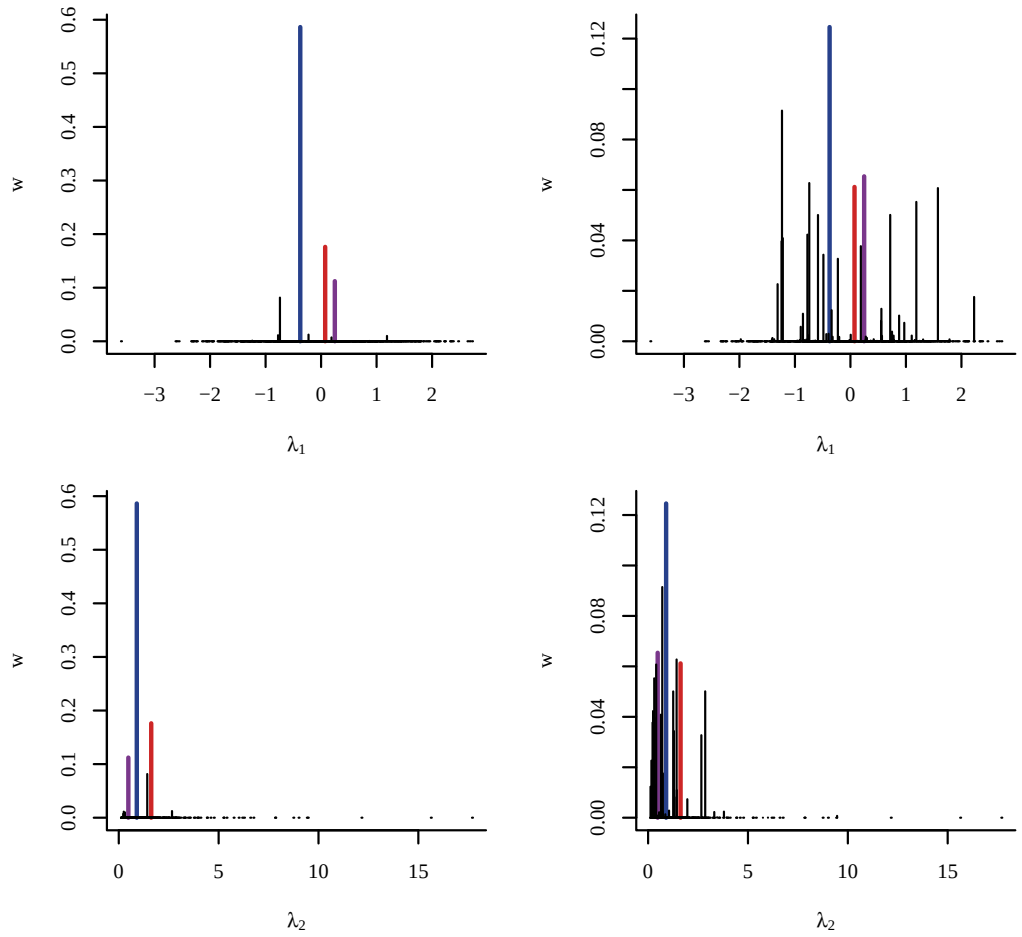

Figure 11 – Representation of the weights computed with the stick-breaking method. The concentration parameter γ equals 1 and 10 for the left-hand and right-hand plots respectively. The baseline distribution is a standard normal in the top row and an inverse-gamma distribution with shape parameter 2 and scale parameter 1 in the bottom row. The same values drawn from the baseline are used for the two different values of γ. The weights w_k, $k = 1,\dots,1000$, were computed in order to sum to 1 and are identical within each column. Blue, violet and red weights represent respectively the first, second and third weights ($k = 1,2,3$) as they came out from the stick-breaking process.

sketch of the model drawn in (4.13) becomes:

$$
\begin{aligned}
X_i \mid c_i, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 &\overset{\text{ind}}{\sim} \mathcal{N}\left(\mu_{c_i}, \sigma^2_{c_i}\right), \quad i = 1,\dots,n, \\
c_i \mid \boldsymbol{w} &\overset{\text{iid}}{\sim} \sum_{k=1}^{N} w_k \delta_k, \quad i = 1,\dots,n, \\
\boldsymbol{w} \mid \gamma &\sim \text{GDirichlet}\left(1,\gamma,\dots,1,\gamma\right), \\
\mu_k \mid \tau &\overset{\text{iid}}{\sim} \mathcal{N}\left(\tau, \sigma^2_{\mu}\right), \quad k = 1,\dots,N, \\
\tau &\sim \mathcal{N}\left(0, \sigma^2_{\tau}\right), \\
\sigma^2_k &\overset{\text{iid}}{\sim} \text{Inv-Gamma}(\nu_1,\nu_2), \quad k = 1,\dots,N, \\
\gamma &\sim \text{Gamma}\left(\eta_1,\eta_2\right),
\end{aligned}
\tag{4.14}
$$

where σ_τ^2 is chosen sufficiently large, and letting η_1 have a large value and $\eta_2 \approx 1/\eta_1$ restricts the concentration parameter prior density to intermediate values for γ, in particular avoiding values too close to 0. Note that the distribution of $\boldsymbol{w} \mid \gamma$ has changed from a Dirichlet distribution in the case where we had a Dirichlet process prior to a generalised Dirichlet distribution GDirichlet$(a_1, b_1, \ldots, a_{N-1}, b_{N-1})$ (Connor and Mosimann, 1969) under the truncation assumption.

We let the blocked Gibbs sampler run for 10,000 iterations and evaluate the posterior distribution obtained on the last 7,000 iterations to get Figure 12. The six components of the mixture are well distinguishable. The pointwise confidence intervals are relatively wide, since they take into account the uncertainty in the mean as well as the uncertainty in the variance. It is interesting to note that the mode of the distribution for the number of components is split between 7 and 8, a slight overestimate of the actual value of 6. As a consequence, we observe overfitting for the galaxy cluster between 30,000 and 35,000 km/s, which is represented with two components in some iterations.

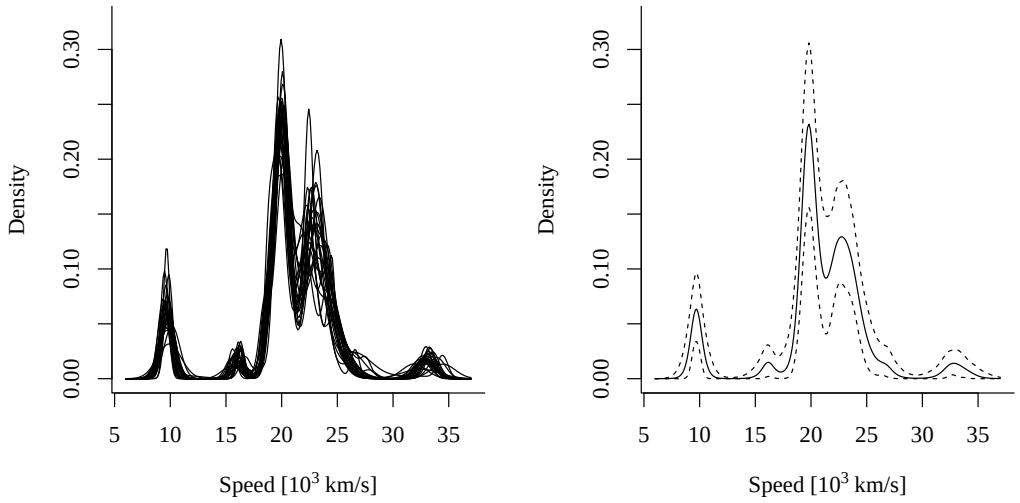

Figure 12 – Posterior density estimation for the galaxy velocity data using the blocked Gibbs sampler. Left panel represents densities from 25 randomly chosen iterations. Right panel is the pointwise mean (solid line) and 2.5% and 97.5% pointwise quantiles (dashed lines) computed on 7,000 iterations.

4.5 Label switching: mixing over components

A common issue arising when dealing with mixture distributions is the problem of *label switching*, first named by Redner and Walker (1984) to describe likelihoods invariant to relabelling. This well-known feature of Markov chain Monte Carlo estimation of mixture models has been addressed in several papers, including Stephens (2000), who describes the multimodality observed on the posterior distributions estimated

on the galaxy data. Jasra *et al.* (2005) review the relabelling methods used in this context, which are all based on deterministic constraints.

In our algorithm however, a permutation of the components results in a different distribution (Porteus *et al.*, 2006). A natural, weak, ordering takes place within the Gibbs sampler, as follows:

$$\mathrm{E}(w_1) \geq \cdots \geq \mathrm{E}(w_N), \tag{4.15}$$

and this may be exploited throughout the sampling process. As we can conclude from the top part of Figure 13, this weak ordering is not observed for the prior weights sampled from the algorithm. There is some probability for $w_k > w_l$, $k < l$, and as we can see in Figure 13 the maximum weight density is unimodal and differs significantly from the density of w_1: the Gibbs sampler can remain stuck in a local minimum, thus not verifying (4.15) any more. Mixing over components is thus needed in order for the sampler to have a better convergence.

The reordering of components has to be stochastic instead of deterministic and the prior weights w_k, $k = 1,\ldots,N$, have to keep their ordering such that their distribution remains unmodified. *Label switching* is thus not to be understood in its exact sense in this framework, since we cannot just swap the indices of two components regardless to their original order. Papaspiliopoulos and Roberts (2008) suggest two complementary label switching moves, which we update to our needs.

The first move consists of swapping two components whose posterior weights disagree with their prior weights. Choose labels k,l in $\{1,\ldots,N\}$ according to the prior distribution of $\boldsymbol{w}$ and ensure that they are both linked to non-empty components. A Metropolis–Hastings acceptance ratio is then computed, which corresponds to permuting n_k with n_l:

$$\frac{w_k^{n_l} w_l^{n_k}}{w_k^{n_k} w_l^{n_l}} \frac{\widetilde{W}_{k+1}^{-1} \cdots \widetilde{W}_l^{-1}}{W_{k+1}^{-1} \cdots W_l^{-1}},$$

if we assume without lost of generality that $k < l$. We write $W_i := \sum_{j=i}^{N} w_j$ and $\widetilde{W}_i := W_i - w_l + w_k$. The case when $l = N$ implies the modified ratio

$$\frac{w_k^{n_N+\gamma-1} w_N^{n_k}}{w_k^{n_k} w_N^{n_N+\gamma-1}} \frac{\widetilde{W}_{k+1}^{-1} \cdots \widetilde{W}_{N-1}^{-1}}{W_{k+1}^{-1} \cdots W_{N-1}^{-1}}.$$

If the move is accepted, the means, variances and posterior weights of the kth and lth components are swapped, and the component indices are reassigned: every $c_i = k$, $i = 1,\ldots,n$, is reset to $c_i = l$ and conversely if $c_i = l$ before the swap, it is set to $c_i = k$ after the swap.

The second move takes place sequentially between neighbour components. For labels k and $k+1$, it proposes to swap the corresponding components in the same sense as for the first move, but to permute V_k and V_{k+1} at the same time. The ratio to be computed is

$$\frac{(1-V_{k+1})^{n_k}}{(1-V_k)^{n_{k+1}}} \frac{\widetilde{W}_2^{-1} \cdots \widetilde{W}_{k+1}^{-1}}{W_2^{-1} \cdots W_{k+1}^{-1}},$$

where now

$$\begin{aligned}
\widetilde{W}_i &:= W_i - w_k - w_{k+1} + \widetilde{w}_k + \widetilde{w}_{k+1}, \quad i = 2,\ldots,k,\\
\widetilde{W}_{k+1} &:= W_{k+1} - w_{k+1} + \widetilde{w}_{k+1},\\
\widetilde{w}_k &:= V_{k+1}\prod_{j=1}^{k-1}\big(1-V_j\big),\\
\widetilde{w}_{k+1} &:= V_k\prod_{j=1}^{k-1}\big(1-V_j\big)(1-V_{k+1}).
\end{aligned}$$

If $k+1 = N$, we have

$$\frac{(1-V_N)^{n_k}}{(1-V_k)^{n_N+\gamma-1}}\,\frac{\widetilde{W}_2^{-1}\cdots\widetilde{W}_{N-1}^{-1}}{W_2^{-1}\cdots W_{N-1}^{-1}}.$$

Figure 13 shows how the label switching procedure leads to a more stable component structure. The almost indistinguishable, multimodal, weight densities in the case where no label switching procedure is used prove the need for jumps across the gaps. Here, the small (82) number of observations makes the gaps between modes smoother and the secondary modes more prominent, but as the dataset increases, the "energy" to cross gaps becomes higher, as noticed by Papaspiliopoulos and Roberts (2008), in which case label switching becomes crucial. Notice also how the density of the first weight is much closer to the density of the maximum weight when adding label switching, which is much more in accordance with the weak ordering (4.15) implied by the stick-breaking representation.

4.6 One-step inference for the Heffernan–Tawn model

Given the blocked Gibbs sampler presented before, we have to include two main features in our model: covariates and multidimensionality. The former is introduced through *dependent Dirichlet processes*, and it can be formulated in terms of the truncated stick-breaking representation as

$$P_{|x} = \sum_{k=1}^{N} w_k \delta_{\boldsymbol{\lambda}_k(x)}, \tag{4.16}$$

so that now a single output of the stick-breaking procedure gives rise to a whole *family* of distributions indexed by x. The data to model are d-dimensional, namely $\boldsymbol{X}_{1,-i},\ldots,\boldsymbol{X}_{n_u,-i}$ given $X_{1,i},\ldots,X_{n_u,i}$ exceed a threshold $u > 0$. We assume them to be a mixture of multivariate normal distributions:

$$\sum_{k=1}^{\infty} w_k \mathcal{MVN}(\boldsymbol{M}_k, \Sigma_k), \tag{4.17}$$

with $M_{k,j} := \alpha_{j|i}x + \mu_{Z,k,j}x^{\beta_{j|i}}$ and $\Sigma_{k,(j,l)} := x^{\beta_{j|i}+\beta_{l|i}}\sigma_{k,(j,l)}$, where $\sigma_{k,(j,l)} := \operatorname{cov}\big(Z_j, Z_l\big)$, the covariance of the Heffernan–Tawn residuals, that is, $\sigma_{k,(j,j)} = \sigma^2_{Z_{j|i},k}$. As before, we

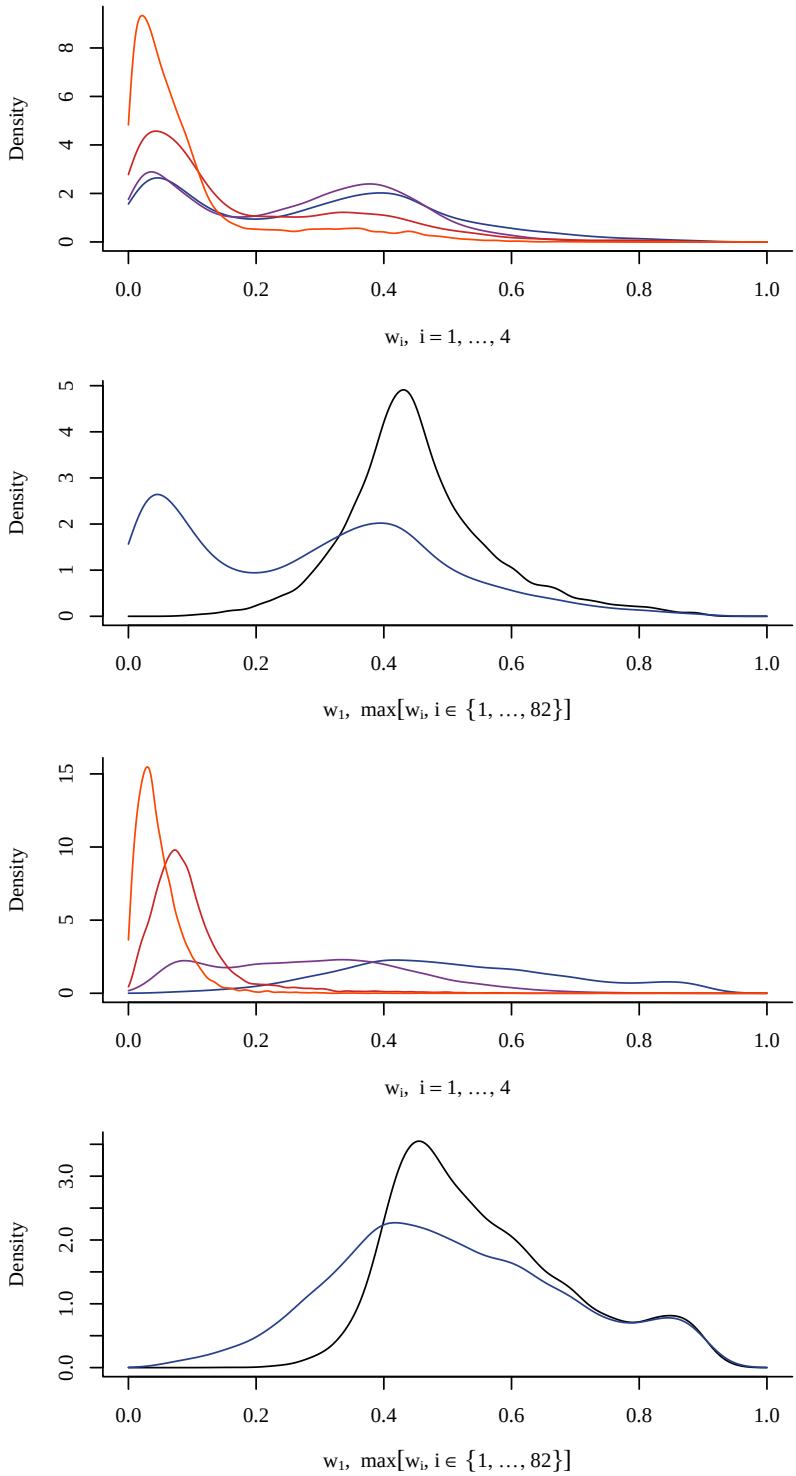

Figure 13 – Comparison of the prior weight densities after a run of the Gibbs sampler on the galaxy data. From the top down the first (without label switching) and third (with label switching) plots show the densities for the first 4 components, in blue, violet, red and orange respectively. The related densities of the first component weights (blue) compared to the densities of the weight maxima (black) are presented in the second and fourth plot from the top down.

use the version of (4.17) truncated at N to allow for sampling directly from posterior distributions.

The introduction of the parametric part has a limited impact on model (4.14) since the hidden variable $\boldsymbol{c}$ allows the stick-breaking representation (4.16) to be split into two parts: the weight sizes w_k on one hand and the weight locations $\boldsymbol{\lambda}_k(x) := \{\boldsymbol{\mu}_k(x), \Sigma_k(x)\}$ on the other hand, $k = 1,\dots,N$. The location part is itself computed from the mean and covariance matrix of the residuals, $\left(\boldsymbol{\mu}_{Z_{|i},k}, \Sigma_{Z_{|i},k}\right)$, with $\Sigma_{Z,k,(j,l)} := \sigma_{k,(j,l)}$, and from the parameters $\boldsymbol{\alpha}_{|i}$ and $\boldsymbol{\beta}_{|i}$.

The final form of our nonparametric model is

$$
\begin{aligned}
\boldsymbol{X}_{l,-i} \mid X_{l,i}, c_l, \boldsymbol{\alpha}_{|i}, \boldsymbol{\beta}_{|i}, \boldsymbol{\mu}_{Z_{|i}}, \Sigma_{Z_{|i}} &\overset{\text{ind}}{\sim} \mathcal{MVN}\left(\boldsymbol{M}_{c_l}, \Sigma_{c_l}\right), \quad X_{l,i} > u,\ l = 1,\dots,n_u, \\
c_l \mid \boldsymbol{w} &\overset{\text{iid}}{\sim} \sum_{k=1}^{N} w_k \delta_k, \quad l = 1,\dots,n_u, \\
\boldsymbol{w} \mid \gamma &\sim \text{GDirichlet}\left(1,\gamma,\dots,1,\gamma\right), \\
\alpha_{j|i} &\overset{\text{iid}}{\sim} \mathcal{U}(0,1), \quad j \neq i,\ j = 1,\dots,d, \\
\beta_{j|i} &\overset{\text{iid}}{\sim} \mathcal{U}(0,1), \quad j \neq i,\ j = 1,\dots,d, \\
\boldsymbol{\mu}_{Z_{|i},k} \mid \boldsymbol{\tau} &\overset{\text{iid}}{\sim} \mathcal{MVN}\left(\boldsymbol{\tau}, \Sigma_{\boldsymbol{\mu}}\right), \quad k = 1,\dots,N, \\
\boldsymbol{\tau} &\sim \mathcal{MVN}\left(\boldsymbol{0}, \Sigma_{\boldsymbol{\tau}}\right), \\
\Sigma_{Z_{|i},k} &\overset{\text{iid}}{\sim} \text{Inv-Wishart}(\nu_1, \boldsymbol{N}_2), \quad k = 1,\dots,N, \\
\gamma &\sim \text{Gamma}\left(\eta_1, \eta_2\right),
\end{aligned}
\tag{4.18}
$$

where the notation is the same as in (4.17), replacing x with $X_{l,i}$, and the inverse Wishart distribution is parametrised with $\nu_1 \in \mathbb{R}$ the degrees of freedom and $\boldsymbol{N}_2$ the symmetric positive definite scale matrix. To allow for more general values for $\boldsymbol{\beta}_{|i}$, its prior density can be set as a normal truncated at 1, or as a uniform joined with a negative exponential

$$
\frac{\lambda_{\beta_{i|j}}}{1+\lambda_{\beta_{i|j}}} \exp\left(\beta_{i|j}\lambda_{\beta_{i|j}}\right) \mathbf{1}\{\beta_{j|i} \leq 0\} + \frac{\lambda_{\beta_{i|j}}}{1+\lambda_{\beta_{i|j}}} \mathbf{1}\{0 \leq \beta_{j|i} \leq 1\},
$$

such that it is possible to tune the density to get more or less mass below 0, by decreasing, respectively increasing, the parameter $\lambda_{\beta_{j|i}} > 0$. The uniform prior is a particular case of this prior, specifically when $\lambda_{\beta_{j|i}} \to \infty$.

Model (4.18) involves a lot of parameters to account for correlations across the dimensions. However the overall dependence is already partly captured through the distribution of the components. We thus assume that observations being part of the same component have independent elements. The covariance matrix Σ_k is now diagonal and the multidimensional parameters $\boldsymbol{\mu}_{Z_{|i},k}$, $\Sigma_{Z_{|i},k} = \text{diag}\left(\sigma^2_{Z_{j|i},k}, j \neq i\right)$, $k = 1,\dots,N$, are sampled elementwise in the same manner as in (4.14):

$$
\begin{aligned}
\mu_{Z_{j|i},k} \mid \tau_j &\overset{\text{ind}}{\sim} \mathcal{N}\left(\tau_j, \sigma^2_{\mu,j}\right), \quad j \neq i,\ j = 1,\dots,d, \\
\sigma^2_{Z_{j|i},k} &\overset{\text{ind}}{\sim} \text{Inv-Gamma}\left(\nu_{1,j}, \nu_{2,j}\right), \quad j \neq i,\ j = 1,\dots,d.
\end{aligned}
$$

As said before, the truncated representation (4.16) allows for calculating the posterior densities in the nonparametric setup. The parameters $\boldsymbol{\alpha}_{|i}$ and $\boldsymbol{\beta}_{|i}$ introduced into the model do not alter this feature, but their prior distributions are not conjugate with respect to a normal likelihood. They have to be sampled through a Metropolis–Hastings step. For the other posterior densities, we give details of their calculation in Appendix B.

The basic scheme followed by the Metropolis–Hastings algorithm is to draw a new value for the parameter we want to estimate from a *proposal distribution* given the previous sampled value of this parameter. It then accepts or rejects this new value based on a ratio which involves the prior density, the likelihood density and the proposal density. In mathematical words, if p denotes the proposal density, q the prior density and l is the likelihood, we have the following procedure to draw a new sample $\theta^{(t)}$ from its posterior, writing $\theta^{(t-1)}$ the value of the parameter θ at the previous step:

- sample $\theta^{(\star)}$ according to $p\left(\cdot \mid \theta^{(t-1)}\right)$,
- compute the *acceptance ratio*

$$\min\left\{1,\ \frac{l\left(X \mid \theta^{(\star)}\right) q\left(\theta^{(\star)}\right) p\left(\theta^{(t-1)} \mid \theta^{(\star)}\right)}{l\left(X \mid \theta^{(t-1)}\right) q\left(\theta^{(t-1)}\right) p\left(\theta^{(\star)} \mid \theta^{(t-1)}\right)}\right\},$$

- accept $\theta^{(\star)}$ as $\theta^{(t)}$ with probability the acceptance ratio; otherwise, set $\theta^{(t)}=\theta^{(t-1)}$.

In our case, the prior density has a bounded support, and so should be the proposal support, in order not to propose infeasible values. On top of that, we know that values 0 and 1 correspond for $\alpha_{j|i}$ and for $\beta_{j|i}$ to specific cases of extremal dependence, so that some mass should be allowed on the bounds. Instead of truncating a normal proposal on $[0,1]$, we could sample $\alpha_{j|i}$ and $\beta_{j|i}$ from

$$\Phi\left(\frac{-\theta^{(t-1)}}{\sigma_\theta}\right)\delta_0 + \left\{\Phi\left(\frac{1-\theta^{(t-1)}}{\sigma_\theta}\right) - \Phi\left(\frac{-\theta^{(t-1)}}{\sigma_\theta}\right)\right\}\mathcal{N}_{(0,1)}\left(\theta^{(t-1)}, \sigma_\theta^2\right) + \left\{1 - \Phi\left(\frac{1-\theta^{(t-1)}}{\sigma_\theta}\right)\right\}\delta_1,$$

where σ_θ^2 is the proposal variance, $\mathcal{N}_{(0,1)}$ specifies a normal distribution truncated to have support on the unit interval and θ can be replaced by $\alpha_{j|i}$ or $\beta_{j|i}$, $j \neq i$, $j = 1, \ldots, d$. In other words, we transfer the mass of a normal distribution below and above $[0,1]$ on 0 and 1 respectively.

However this proposal density cannot be directly implemented since the acceptance ratio would not be well-defined. We use instead an informative beta prior to mimic Dirac masses on the boundaries. The proposal can then be a normal truncated on the unit interval, appearing in the acceptance ratio only through its normalising constant. See Appendix C for a discussion on alternative methods to sample from the posterior distributions of $\boldsymbol{\alpha}_{|i}$ and $\boldsymbol{\beta}_{|i}$.

5 Comparative study: two- and one-step procedures

To assess the efficiency of the Bayesian semiparametric method, we compare it with the two-step method on simulated data. The peakflow data allow for an application of the one-step method to the hydrological context.

5.1 Simulated data

In order to have a benchmark on which to rely for both methods, we generate data from a Gumbel copula, as suggested by Keef *et al.* (2009). The Gumbel dependence structure is

$$\exp\left[-\left\{\sum_{i=1}^{d}\exp(-x_i/\zeta)\right\}^{\zeta}\right], \quad \zeta \in (0,1],\ \boldsymbol{x} \in \mathbb{R}^d,$$

for which independence is reached when $\zeta = 1$, and $0 < \zeta < 1$ corresponds to asymptotic dependence. In the latter case we have already seen in §3.1 that the Heffernan–Tawn model parameters $\boldsymbol{\alpha}_{|i}$ and $\boldsymbol{\beta}_{|i}$, $i = 1,\ldots,d$, take values 1 and 0 respectively.

We choose two different values for the copula dependence parameter, namely $\zeta = 0.2$ and $\zeta = 0.8$, and generate 12,000 data triplets for each case. After having marginally transformed the sample to the Gumbel scale, we select the 5% largest data on one margin and retain the corresponding data in the two other margins, so that the inference is made on 600 data triplets, which is approximately the amount of observations we shall consider in the next section. We repeat this on 200 samples and compare the two- with the one-step method.

For the one-step method, we fix the maximum number of components to $N = 50$ and let the algorithm run for 6,000 iterations following a burn-in of 4,000 iterations. For each generated dataset, we compute the mean and the median of $\widehat{\boldsymbol{\alpha}}_{|1}$ and $\widehat{\boldsymbol{\beta}}_{|1}$, as the latter is more robust for estimates close to the boundaries. The marginal residual densities are summarised through their pointwise mean for the 200 simulated datasets.

As the dependence structure is the same between each margin of the Gumbel copula, we present the results only for one of them. Figure 14 shows two mirror histograms meant to compare the distribution of $\widehat{\alpha}_{3|1}$ and $\widehat{\beta}_{3|1}$ on the 200 datasets generated from a Gumbel copula with $\zeta = 0.2$. The parameter estimates $\widehat{\alpha}_{3|1}$ are much more spread across the unit interval in the case of the one-step method. More than 90% of these estimates are larger than 0.9, and the flexibility offered by the Dirichlet process tends to allow for estimates far from the true value by changing the residual density accordingly. Observe the accuracy of the Bayesian method in estimating $\beta_{3|1}$, with a lot of mass near 0, and some outliers trying to balance outliers in $\widehat{\alpha}_{3|1}$. The corresponding estimators given by the two-step procedure are all in the bottom part of the unit interval, but there is less mass near 0.

The case where $\zeta = 0.8$, as it is closer to independence, gives estimators which reflect the case of extremal dependence with asymptotic independence. The two methods give similar results, not shown here, with the Bayesian approach offering much

more flexibility; the parameters and residual density of the Heffernan–Tawn model compensate each other and can individually be far from their true values.

5.2 River peakflows

In this application, we focus on the peakflows based on the observations collected in the six gauging stations from south England (cf. §1.1). Taking the river Thames as the conditioning variable seems reasonable, since the five other rivers are its — direct or indirect — tributaries, so that the Thames might explain a major part of the dependency between the series. We assume that peakflows, by definition, are propagated sufficiently quickly from the tributaries to the river Thames such that high levels of water happen on the same day in the six different places.

In the context of real data, we first have to transform the marginal distributions to the Gumbel scale, using the semiparametric estimation (1.4). The choice of each of the six thresholds is made separately on the six series, and six generalised Pareto distributions are fitted to the resulting exceedances. For the choice of the conditioning threshold, we estimate the Heffernan–Tawn model at different levels of conditioning and seek for the lowest level such that $\widehat{\boldsymbol{\alpha}}_{|1}$ and $\widehat{\boldsymbol{\beta}}_{|1}$ remain stable above it. The plots for this exploratory analysis are shown in Figure 15. It has been computed for 10 different levels, on the last $7,000$ iterations of a $10,000$ iteration run, and with a maximum number of components set to 100. We select the 97% threshold, as we know that $\widehat{\boldsymbol{\beta}}_{|1}$ is difficult to estimate and thus has more erratic shapes than $\widehat{\boldsymbol{\alpha}}_{|1}$. We let the Gibbs sampler run through $25,000$ iterations and skip the first $5,000$, and set the maximum number of components to 150, in order to have a stick-breaking procedure almost indistinguishable from a non-truncated Dirichlet process.

We show here only the main features of this output, and the reader interested in more details can refer to Appendix D. Figure 16 compares a marginal residual density obtained with the two- and one-step methods. The same kind of shapes are found for the other marginal residuals. To show the goodness-of-fit for the full peakflow distribution, Figure 17 presents a comparison between the joint density of (X_2, X_3) conditioned on X_1 and the corresponding joint density estimated from the Ray's and Lambourn's peakflows conditioned on the Thames's high peakflow. Contours are based on kernel density estimates based on data simulated from the Gibbs sampler output, and on the real data, using a rule-of-thumb for the choice of the bandwidth (Venables and Ripley, 2002). The overall shape is well captured, confirming the quality of the general fit.

We conclude that the one-step method gives fits of the same quality as the two-step method, with the advantage of giving more insight into the uncertainty of the residual distribution and, more importantly, into the uncertainty of the whole conditional distribution.

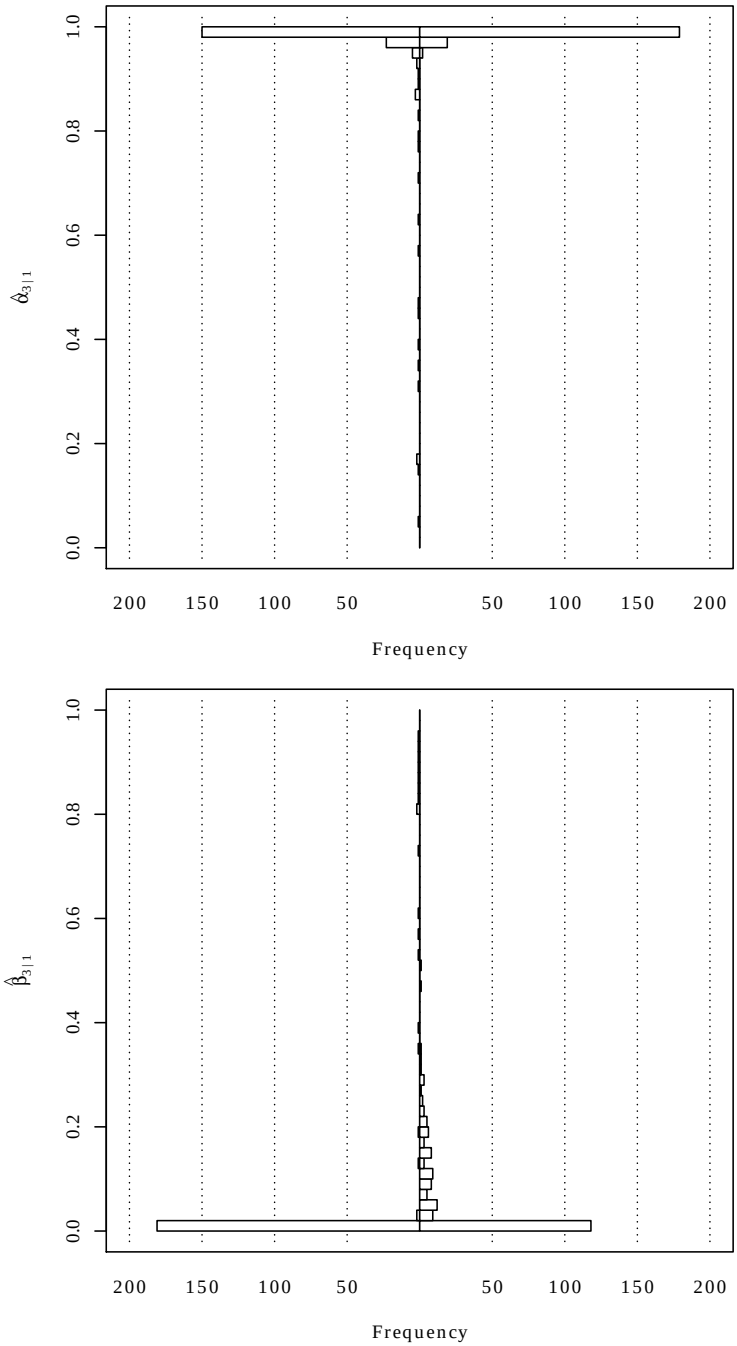

Figure 14 – Histograms for the Heffernan–Tawn parameter estimators fitted on 200 samples from a Gumbel copula with dependence parameter $\zeta = 0.2$. The left-hand side histograms stand for the medians of the Gibbs sampler output and the right-hand side histograms are the corresponding frequency of $\widehat{\alpha}_{3|1}$ and $\widehat{\beta}_{3|1}$ based on the two-step approach.

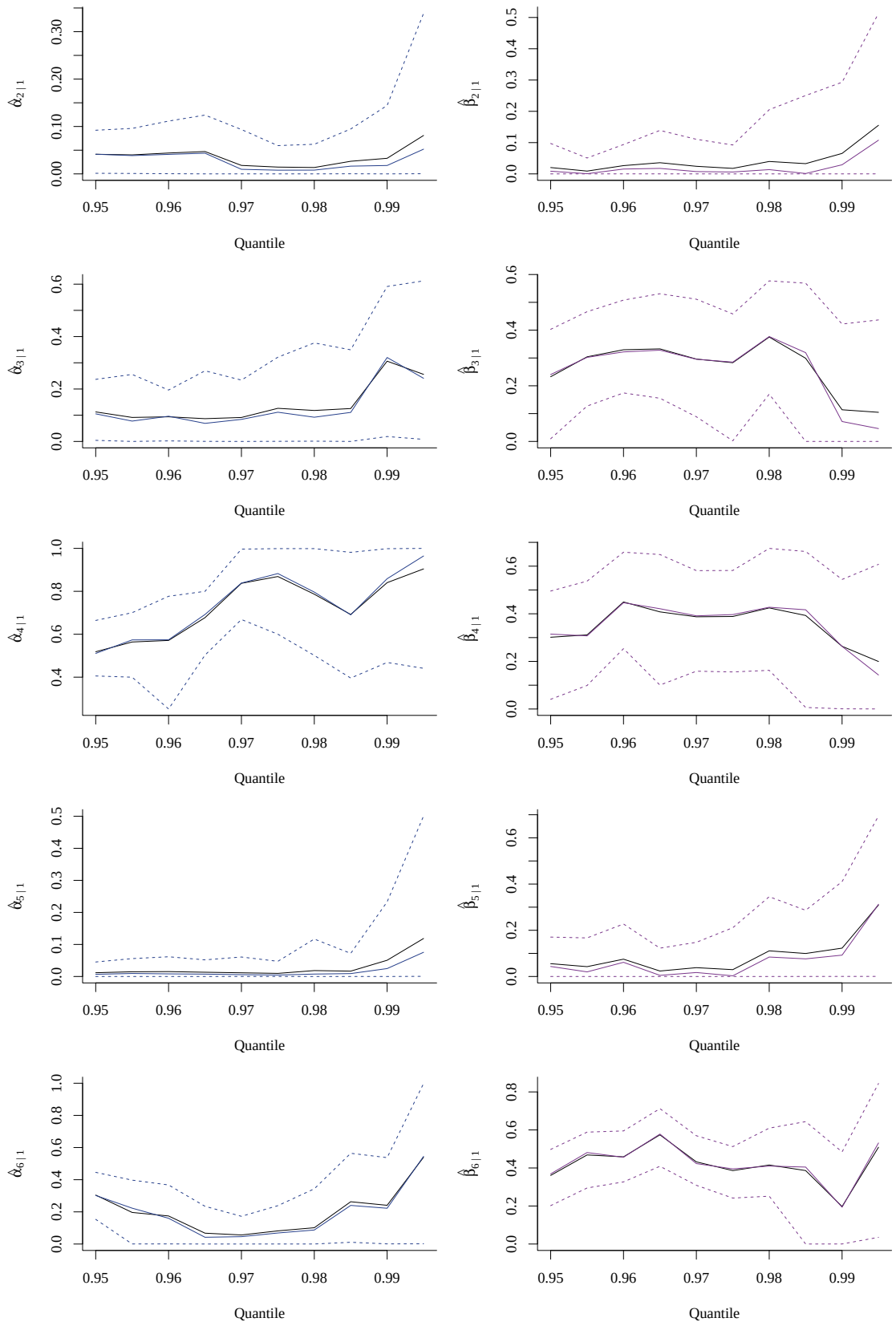

Figure 15 – Heffernan–Tawn model fitted on the peakflow data for different conditioning thresholds ranging from the 95% quantile up to the 99.5% quantile. Solid black lines represent the means on 7,000 iterations following a 3,000 iteration burn-in, and coloured lines correspond to the medians (solid) and to the 2.5% and 97.5% quantiles (dashed) computed on the same 10 Gibbs sampler outputs.

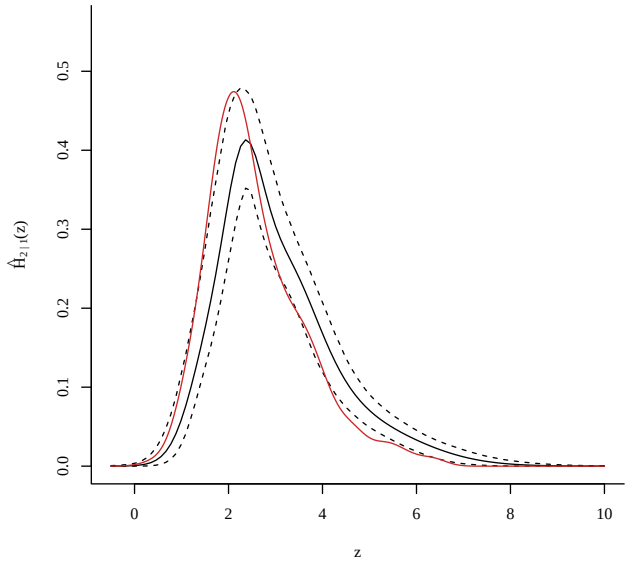

Figure 16 – Density of Heffernan–Tawn residuals corresponding to the Ray's peakflow conditioned on high levels of the Thames' peakflow. The black lines show the pointwise mean (solid line) and pointwise 2.5% and 97.5% quantiles (dashed) computed on 2,000 iterations randomly chosen from the output of the Gibbs sampler. The solid red line corresponds to a kernel density estimate of the residuals based on a two-step fit.

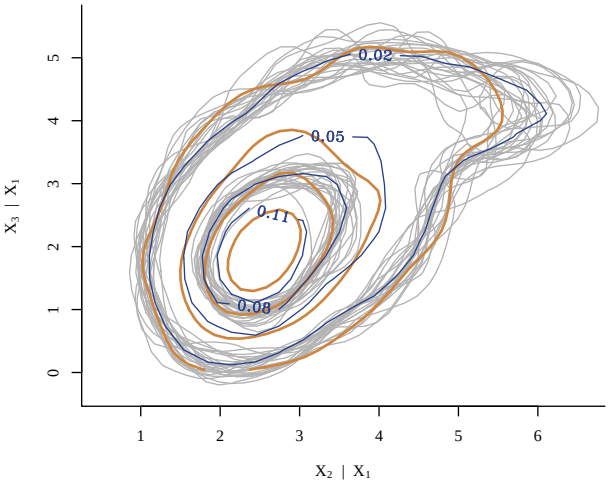

Figure 17 – Contours of the joint density of (X_2, X_3) conditioned on high levels of X_1. Grey lines are contour lines for 25 randomly chosen iterations from the Gibbs sampler output, corresponding to densities of 0.02 and 0.08. Each one is based on 1,500 simulated data. Blue contours come from the kernel density estimated on 300 such iterations. The related contours computed from the original data are shown in light brown.

Distribution of cluster maxima

We have already seen (cf. beginning of §2.2) that short-range dependence within extreme values can be modelled by approximating the mean cluster length θ^{-1} and using this information to modify the survivor distribution of excesses or by considering only cluster maxima to make the inference. These models are based on asymptotic results and are only approximations at a high threshold u. The former assumes that θ is stable above u, but as already pointed out by Smith and Weissman (1994) we shall see that this assumption is too rough in most cases. The latter introduces a significant bias partly due to the arbitrary cluster selection and is liable not to be robust against changes in the run-length m. We describe a novel, subasymptotic approach, followed by a comparison with the peaks over threshold approach on simulated data, and conclude with an application on the peakflow series.

6 Subasymptotic model for the cluster maxima

Eastoe and Tawn (2012) considered a subasymptotic model which accounts for variation in θ at high quantiles. Consider the subasymptotic extremal index

$$\theta(x,m) := \Pr(X_2 \leq x, \ldots, X_m \leq x \mid X_1 > x), \tag{6.1}$$

where m is such that two clusters are at least $m-1$ observations away from each other. Their idea is to modify the conditional survivor distribution in (2.10) to get the cluster maxima conditional survivor distribution:

$$\frac{\theta(x,m)}{\theta(u,m)}\left(1+\xi\frac{x-u}{\sigma_u}\right)_+^{-1/\xi}, \quad x > u. \tag{6.2}$$

A simple argument can be used to justify expression (6.2) as the conditional survivor distribution for the cluster maxima. Using the interpretation of the extremal index as the proportion of peaks (written X_{clust}) amongst all excesses (X_{all}), we get:

$$\begin{aligned}\frac{\theta(x,m)}{\theta(u,m)}\Pr(X_{\text{all}} > x \mid X_{\text{all}} > u) &= \frac{\Pr(X_{\text{clust}} > x)}{\Pr(X_{\text{all}} > x)}\,\frac{\Pr(X_{\text{all}} > u)}{\Pr(X_{\text{clust}} > u)}\,\frac{\Pr(X_{\text{all}} > x)}{\Pr(X_{\text{all}} > u)} \\ &= \frac{\Pr(X_{\text{clust}} > x)}{\Pr(X_{\text{clust}} > u)} \\ &= \Pr(X_{\text{clust}} > x \mid X_{\text{clust}} > u),\end{aligned}$$

where the last probability is the survivor distribution for the peaks appearing beyond the threshold u.

6.1 Inference for the subasymptotic extremal index

To make inference about the subasymptotic extremal index defined in (6.1), we observe that the formulation of the Heffernan–Tawn model exactly matches our needs.

From the asymptotic conditional independence (3.4) we can derive $\theta(x,m)$ for large x in the following way:

$$\begin{aligned}\theta(x,m) &:= \Pr(X_2 \le x,\dots,X_m \mid X_1 > x) = \\ &= \frac{\Pr(X_1 > x, X_2 \le x,\dots,X_m \le x)}{\Pr(X_1 > x)} \\ &= \frac{\int_x^\infty \Pr(X_1 = y, X_2 \le x,\dots,X_m \le x)\,\mathrm{d}y}{\Pr(X_1 > x)} \\ &= \int_x^\infty \Pr\left(X_1 = y, \boldsymbol{Z}_{|1}(y) \le \boldsymbol{z}(x,y) \middle| X_1 > x\right)\mathrm{d}y \\ &\approx \int_x^\infty g_x(y) H_{|1}\{\boldsymbol{z}(x,y)\}\,\mathrm{d}y, \quad x > u, \end{aligned} \tag{6.3}$$

where u is a suitable threshold for the Heffernan–Tawn model, $g_x(y)$ is the density of a generalised Pareto distribution for threshold x, $\boldsymbol{Z}_{|1}(y)$ is calculated based on (3.3) and

$$z_j(x,y) = \frac{T_{\mathscr{G}}(x) - \alpha_{j|1} T_{\mathscr{G}}(y)}{T_{\mathscr{G}}(y)^{\beta_{j|1}}} \tag{6.4}$$

is the element of $\boldsymbol{z}(x,y)$ associated with X_j.

Now that we have a model for $\theta(x,m)$ we can apply both the two-step and the one-step methods in §3.2 and §4.6 respectively to get estimators for $\boldsymbol{\alpha}_{|1}$, $\boldsymbol{\beta}_{|1}$ and for the distribution of $\boldsymbol{Z}_{|1}$. We first present the inference described by Eastoe and Tawn (2012) to then show how it can be updated to the features of the one-step method.

6.1.1 Using the two-step method

Within the framework of the two-step method, we have a sample of residuals with dimension $(m-1)$, viz. $\widehat{\boldsymbol{Z}}_{|1}(X_{1,1}),\dots,\widehat{\boldsymbol{Z}}_{|1}(X_{1,n_u})$.

Method of proportion Sample R of those residuals with replacement and generate R replicates $X_1^{(1)},\dots,X_1^{(R)}$ from a generalised Pareto distribution. Compute $\boldsymbol{X}_{-1}^{(r)}$, as

$$T_{\mathscr{G}}\left(X_j^{(r)}\right) = \widehat{\alpha}_{j|1} T_{\mathscr{G}}\left(X_1^{(r)}\right) + T_{\mathscr{G}}\left(X_1^{(r)}\right)^{\widehat{\beta}_{j|1}} Z_{j|1}^{(r)}, \quad j = 2,\dots,m, \quad r = 1,\dots,R$$

and use the inverse of transformation $T_{\mathscr{G}}$ on the left-hand side. Eastoe and Tawn (2012) propose to use the same samples from $\widehat{\boldsymbol{Z}}_{|1}$ and from the generalised Pareto distribution for all values of x to reduce Monte Carlo variation.

Monte Carlo integration The other method to derive an estimator for $\theta(x,m)$ involves Monte Carlo integration. After sampling X_1 in the same way as above, compute $z_j^{(r)} := z_j(x, X_1^{(r)})$, $j = 2,\dots,m$, using (6.4) and evaluate the empirical distribution function

$$\widehat{H}_{|1}^{(r)} = \widehat{H}_{|1}\left(z_2^{(r)},\dots,z_m^{(r)}\right), \quad r = 1,\dots,R.$$

Finally compute the mean of evaluations $\widehat{H}_{|1}^{(1)},\ldots,\widehat{H}_{|1}^{(R)}$ to get an approximation to integral (6.3) of the form

$$\widehat{\theta}(x,m) = \frac{1}{R}\sum_{r=1}^{R}\widehat{H}_{|1}^{(r)}.$$

Confidence intervals for $\widehat{\theta}(x,m)$ do not really bring more information since they are simply given by a binomial distribution which does not take into account the uncertainty in $\widehat{\boldsymbol{\alpha}}_{|1}$ and $\widehat{\boldsymbol{\beta}}_{|1}$. Eastoe and Tawn (2012) used a bootstrap method to get confidence bounds. The benefit that we get using the one-step method is that confidence intervals are directly deduced from the fit.

6.1.2 Using the one-step method

A one-step fit on the d-dimensional observations $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{n_u}$, $X_{i,1} > u$, $i = 1,\ldots,n_u$, is characterised by the maximum number of components N for the truncated stick-breaking representation, the number of iterations S and the following parameters for each iteration:

- $\widehat{\boldsymbol{\alpha}}_{|1}$, $\widehat{\boldsymbol{\beta}}_{|1}$ $(d-1)$-dimensional vectors,
- $\widehat{\boldsymbol{\mu}}_Z$, $\widehat{\boldsymbol{\sigma}}_Z^2$ matrices with dimensions $N\times(d-1)$,

plus some additional information giving the mapping between the observations and their corresponding component at each iteration, of the form c_i, $i = 1,\ldots,n_u$. Associated to each component are the weights $\boldsymbol{w}$, specified for each iteration, but common to each dimension.

In this one-step framework the distribution $H_{|1}$ of $\boldsymbol{Z}_{|1}$ involves some more complexity, specifically

$$H_{|1}\{z_2(x,y),\ldots,z_d(x,y)\} = \sum_{k=1}^{N} w_k \prod_{j=2}^{m}\Phi\left\{\frac{z_j(x,y)-\mu_{Z,k}}{\sigma_{Z,k}}\right\},$$

or in terms of the effective component weights, i.e., by considering the number of observations per component,

$$H_{|1}\{z_2(x,y),\ldots,z_d(x,y)\} = \frac{1}{n_u}\sum_{i=1}^{n_u}\prod_{j=2}^{m}\Phi\left\{\frac{z_j(x,y)-\mu_{Z,c_i}}{\sigma_{Z,c_i}}\right\}. \tag{6.5}$$

Method of proportion The method can be outlined as follows: generate independent observations from the generalised Pareto distribution and draw independent residuals $\boldsymbol{Z}_{|1}$ from $\widehat{H}_{|1}$, the function estimate of (6.5). Those samples are substituted in the Heffernan-Tawn model and give $\boldsymbol{X}_{-1}$. This procedure has to take place through S iterations in order to reflect the variation in $\widehat{\boldsymbol{\alpha}}_{|1}$ and $\widehat{\boldsymbol{\beta}}_{|1}$, as well as the uncertainty in $\widehat{H}_{|1}$.

Given the information above we are able to generate a new sample of $\boldsymbol{X}$, namely $\boldsymbol{X}^{(1)},\ldots,\boldsymbol{X}^{(R)}$, with $X_1^{(r)} > x$, $r = 1,\ldots,R$, in the following way: assume that we

want to estimate $\theta(x,m)$ for a range of discrete x-values within $[x_{\text{bottom}}, x_{\text{up}}]$ and $m = 2,\ldots,d$. We first simulate R standard uniform variables $X_U^{(1)},\ldots,X_U^{(R)}$, which will then be transformed into a Gumbel sample for each different value of x_G, i.e., for each value of x in Gumbel scale. Before looping on each x value, we also draw R $(d-1)$-dimensional values for the residuals $\boldsymbol{Z}_{|1}$ from distribution $H_{|1}$ in (6.5):

(1) for each iteration s in $\{1,\ldots,S\}$ and each sample index r in $\{1,\ldots,R\}$, randomly pick a datum index i in $\{1,\ldots,n_u\}$;

(2) for this i, at one specific iteration, we know the corresponding $\widehat{\mu}_{Z,c_i,j}$, $\widehat{\sigma}^2_{Z,c_i,j}$, $j = 2,\ldots,d$, of the corresponding multivariate normal component;

(3) we can now generate $Z_{j|1}^{(r,s)}$ as a normal with the mean and the variance given in the previous point, and we repeat this procedure for $(s,r) \in \{1,\ldots,S\}\times\{1,\ldots,R\}$.

At the end of the process, we have generated $(d-1)RS$ values for the residuals. Notice that we do not use the component weights $w_1,\ldots,w_N$ computed by the Gibbs sampler and rather choose the effective weights $n_1,\ldots,n_N$ which represent the number of observations per component or, in the Bayesian formulation, we use the posterior distribution of $\boldsymbol{w}$.

After these preliminaries, we come to the computation of $\widehat{\theta}(x_G,m)$ itself. For each value of x_G

(1) define the conditioning variable in Gumbel scale $X_{G,1}^{(r)}$ as

$$-\log\left[-\log\left\{x_G+(1-x_G)X_U^{(r)}\right\}\right], \quad r = 1,\ldots,R;$$

(2) compute the $(d-1)$-dimensional conditioned variable $\boldsymbol{X}_{G,-1}^{(r,s)}$ as

$$X_{G,1}^{(r)}\widehat{\boldsymbol{\alpha}}_{|1}^{s}+\exp\left\{\log\left(X_{G,1}^{(r)}\right)\widehat{\boldsymbol{\beta}}_{|1}^{s}\right\} * \boldsymbol{Z}_{|1}^{(r,s)}, \quad r = 1,\ldots,R,\ s = 1,\ldots,S,$$

where $*$ is the *Hadamard product*, i.e., the componentwise vector multiplication. The matrix product can help to avoid several levels of nested loops in a very powerful way (cf. Appendix E);

(3) extract the maximum value $M_m^{(r,s)}$ across dimensions 2 to m, with $m = 2,\ldots,d$, of $\boldsymbol{X}_{G,-1}^{(r,s)}$ for each r in $\{1,\ldots,R\}$ and each iteration s in $\{1,\ldots,S\}$;

(4) for each iteration s and each m compute the proportion

$$\sharp\left\{r \in \{1,\ldots,R\} : M_m^{(r,s)} < x_G\right\}\Big/R,$$

which gives $\widehat{\theta}^{(s)}(x_G,m)$ for each $m = 2,\ldots,d$ on each of the S iterations;

(5) the pointwise confidence intervals are the quantiles estimated across the S values of $\widehat{\theta}(x_G,m)$, and this can be done for all m in $\{2,\ldots,d\}$.

Observe that we use the same sample of conditioning variables $X_{G,1}^{(r)}$, the same set of randomly-generated residuals $\boldsymbol{Z}_{|1}^{(r,s)}$ and the same normalising parameters $\widehat{\boldsymbol{\alpha}}_{|1}^{s}$ and $\widehat{\boldsymbol{\beta}}_{|1}^{s}$ for each different value of x_G to reduce the Monte Carlo variation, based on the idea of Eastoe and Tawn (2012).

Monte Carlo integration We generate the conditioning variable as presented for the beginning of the proportion method. The idea is to compute the residuals $\boldsymbol{Z}_{|1}$ based on the model, through all dimensions, sampled values and iterations. We then evaluate the estimated version $\widehat{H}_{|1}$ of (6.5) on those residuals. Monte Carlo integral estimates of $\theta(x,m)$ can eventually be computed for each iteration on samples of size R.

For each value of x_G

(1) compute $X_{G,1}^{(r)}$, r=1,...,R as presented within the proportion method framework;

(2) draw values for the residuals

$$Z_{j|1}^{(r,s)} = \frac{x_G - \widehat{\alpha}_{j|1}^{s} X_{G,1}^{(r)}}{\exp\left\{\widehat{\beta}_{j|1}^{s} \log\left(X_j^{(r,s)}\right)\right\}}, \quad j = 2,\ldots,d,\ s = 1,\ldots,S,\ r = 1,\ldots,R,$$

(3) evaluate $\widehat{H}_{|1}(z_2,\ldots,z_d)$ on those residuals: for each iteration $s \in \{1,\ldots,S\}$ and each $r \in \{1,\ldots,R\}$, add up the (weighted) probabilities of each Gaussian component:

$$\widehat{H}_{|1}^{(r,s)} = \frac{1}{n_u} \sum_{i=1}^{n_u} \prod_{j=2}^{m} \Phi\left\{ \frac{Z_{j|1}^{(r,s)} - \widehat{\mu}_{Z,c_i}^{s}}{\widehat{\sigma}_{Z,c_i}^{s}} \right\}, \quad r = 1,\ldots,R \quad s = 1,\ldots,S,$$

where again the weights are computed in a way such that only the effective size of components matters;

(4) take the mean across the S samples of size R obtained in previous step to get a set of estimates $\widehat{\theta}^{(s)}(x_G,m)$, $s = 1,\ldots,S$;

(5) the mean over those S estimates gives $\widehat{\theta}(x_G,m)$ and confidence intervals for this estimator are given by the empirical quantiles computed on $\widehat{\theta}^{(s)}(x_G,m)$, $s = 1,\ldots,S$.

6.2 Inference for the distribution of cluster maxima

Using the POT approach to estimate the distribution of cluster maxima involves a new fit for every different value of the run-length m: the fit depends on the actual values of the cluster maxima, selected after having determined the clusters according to m. This can be cumbersome and leads to highly varying estimates in the distribution

$$\Pr(X_{\text{clust}} \leq x \mid X_{\text{clust}} > u) = 1 - \left(1 + \xi_{\text{clust}} \frac{x-u}{\sigma_{\text{clust}}}\right)^{-1/\xi_{\text{clust}}}, \quad x > u,$$

where X_{clust} is a random variable distributed as a cluster maximum and ξ_{clust} and σ_{clust} have estimators which only depend on the cluster maxima.

One of the advantages of the subasymptotic model is that it splits the estimation between the marginal distribution of exceedances on one hand and the clustering effect on the other hand, namely

$$\Pr(X_{\text{clust}} \leq x \mid X_{\text{clust}} > u) = 1 - \frac{\theta(x,m)}{\theta(u,m)} \left(1 + \xi_{\text{all}} \frac{x-u}{\sigma_{\text{all}}}\right)^{-1/\xi_{\text{all}}}, \quad x > u, \tag{6.6}$$

with ξ_{all} and σ_{all} indicating that their respective estimators are based on all exceedances of u. As a consequence of formulation (6.6), only the extremal index ratio changes for different values of m. More than that, the Heffernan–Tawn fit for $m-1$ can be reused for m when using the two-step method: thanks to the working assumption of conditional independence of the residuals, we can just fit the model for lag m, i.e., for X_m given $X_1 > u$. From there we get $\widehat{\alpha}_{m|1}$, $\widehat{\beta}_{m|1}$ and the empirical distribution of the estimated residuals $\widehat{\boldsymbol{Z}}_{m|1}$. For the one-step method, anticipation is needed, and the fit for the maximum run-length $m_{\max}$ has to be computed in the first place, so that subsets of the output can be used to estimate the ratio $\theta(x,m)/\theta(u,m)$ for $m \leq m_{\max}$.

Using the method described in the previous section to estimate $\theta(x,m)$ after a one-step fit of the Heffernan–Tawn model, it is easy to compute the estimator for the ratio $\theta(x,m)/\theta(u,m)$ by taking the ratio for each of the S iterations, i.e., $\widehat{\theta}^{(s)}(x,m)/\widehat{\theta}^{(s)}(u,m)$, and then computing their mean and appropriate quantiles to get an estimator and confidence bounds. If the Heffernan–Tawn model is fitted with the two-step method, then the estimator for the ratio is simply the ratio of the estimators $\widehat{\theta}(x,m)/\widehat{\theta}(u,m)$.

7 Comparative study: POT and subasymptotic model

We now present the performance of the peaks over threshold compared to the subasymptotic approach. To simplify the discussion and lighten the plots, we present the latter only when fitted with the one-step method, as the results are similar to the ones obtained with the two-step method, with the advantage of providing confidence intervals directly. In a second part, we show an example involving Lambourn peakflow data.

7.1 Simulated data

The simulated data correspond to process 1 described by Eastoe and Tawn (2012), and has the structure of dependence of the bivariate Gaussian copula, with dependence parameter ρ varying in $(-1,1)$. This process is asymptotically independent, corresponding to the extremal index $\theta = 1$. It is set with exponentially distributed margins, so that we do not introduce an error through the choice of the marginal thresholds.

We fitted the peaks over threshold and the subasymptotic models on 102 datasets generated with such a process, each of size $6,000$. The marginal threshold u was fixed at the 90% empirical quantile of each series, and the run-length m ranged from 2 to 12. We then estimated several high quantiles of the distribution of cluster maxima conditioned on u. The true distribution of cluster maxima is approached with the empirical distribution of cluster maxima computed on a basis of 10^6 observations generated from the same process.

Figure 18 compares the 95% and 99.9% quantiles estimated with the peaks over threshold method and with the subasymptotic model using the one-step method. We use boxplots to better reflect the distribution of estimates across the generated datasets. The approximated true distribution stands as a benchmark. The estimates given by the peaks over threshold approach are a lot more variable than the ones from the subasymptotic model. They also depend much more on the value of m, as the whole distribution needs to be estimated for each different cluster definition. The already wide divergence observed between the two models at the 95% level becomes huge at the 99.9% level. Not only does the stability of the estimator for the subasymptotic model across different values of m appear clearly, but it is also much less biased than the peaks over threshold quantile.

As a complement to this figure, we list in Table 3 the mean square error for two extremal values of m at different quantiles, for the peaks over threshold approach and for the subasymptotic model, estimated with both the two- and one-step method. As mentioned in §5.1 when dealing with the Heffernan–Tawn model, the individual parameters can differ between the one- and the two-step methods, but the overall distribution is very similar. This is partly shown by the similar figures in Table 3.

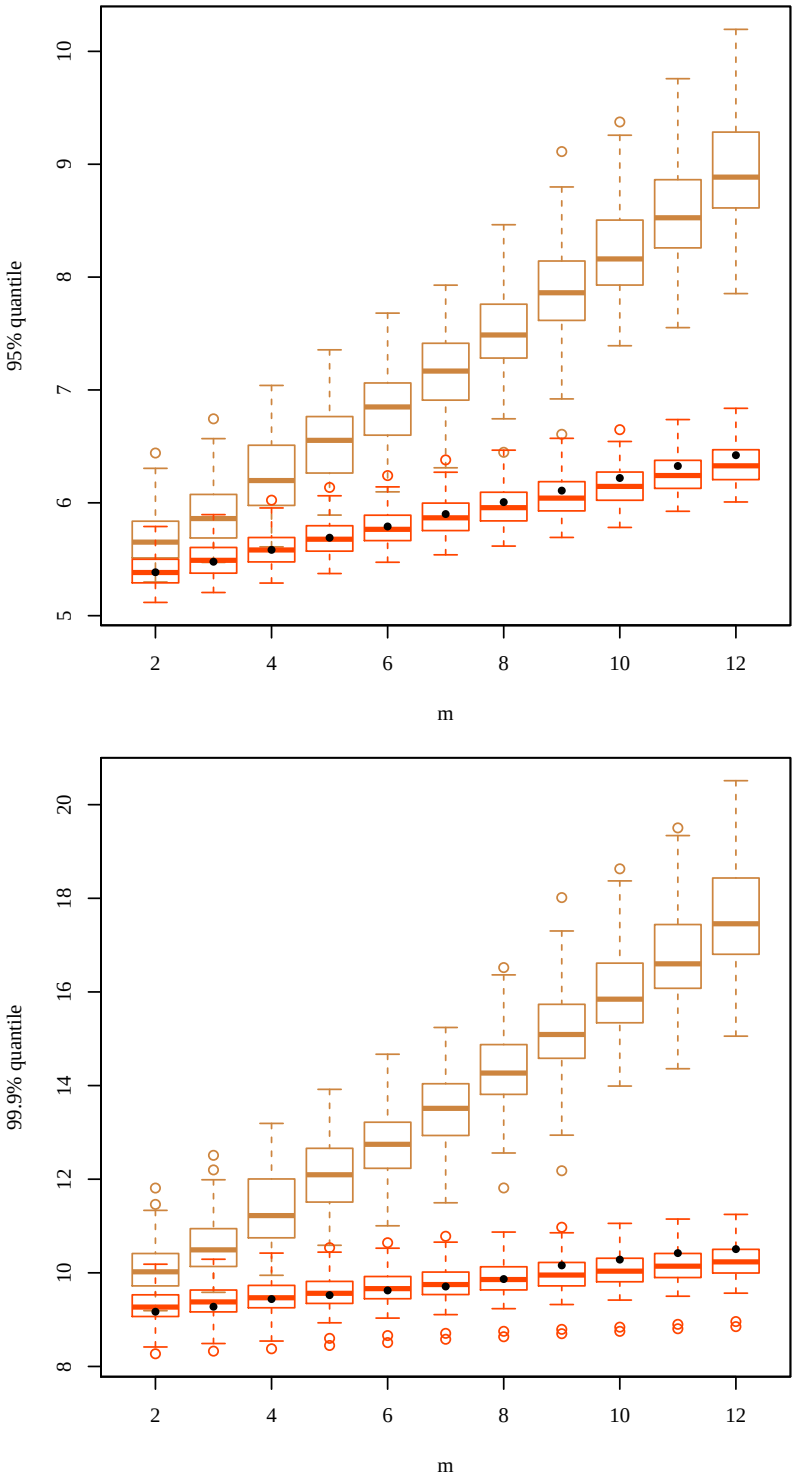

Figure 18 – Quantiles for the conditional cluster maxima distribution estimated with the peaks over threshold method (light brown) and with the subasymptotic method (blue). The true distribution estimated on a sample of size 10^6 is shown in black. Top panel: 95% quantile; bottom panel: 99.9% quantile.

	$m = 2$			$m = 12$		
Quantile	POT	SUB	BAY	POT	SUB	BAY
95%	10	.01	.008	700	.002	.6
98%	20	.03	.03	1000	.001	.5
99%	30	.02	.03	2000	.04	.3
99.9%	100	1	1	5000	4	7

Table 3 – Estimated mean square error on 102 simulated datasets. POT: peaks over threshold; SUB: subasymptotic model fitted with the two-step method; BAY: subasymptotic model fitted with the one-step method.

7.2 River peakflow

River peakflows are typically short-range dependent: after an important rainfall the peakflows can reach high values, and decrease gradually, while the river gets back into the low flow regime. We thus expect peakflows to be dependent at extreme levels and at small lags m. We choose the Lambourn's peakflow to carry out this analysis on real data.

A base comparison available at low levels is the runs estimator (2.9), and block bootstrap can give appropriate confidence intervals. The advantage of parametric modelling is that we can extrapolate the subasymptotic extremal index beyond the largest observation. We only show $\theta(u,m)$ estimated with the one-step method (Figure 19), since it provides natural confidence intervals and also because the two-step method gives similar results.

The choice of m is based on exploratory fits of the Heffernan–Tawn model. The decrease of $\widehat{\alpha}_{m|1}$ and $\widehat{\beta}_{m|1}$ for increasing m in Figure 20 indicates that independence is reached from $m = 12$ on, value for which the two parameters vanish. The decrease in $\alpha_{m|1}$ appears close to a geometric series, while $\beta_{m|1}$ seems constant or linear in m, suggesting that a parametric structure could be used to model these parameters across a range of lags. The values of $\widehat{\beta}_{6|1}$ and $\widehat{\beta}_{10|1}$ illustrate an issue regarding the beta prior; if the Markov chain gets trapped on a value close to 0, the probability for a Markov chain move becomes very small, leading to a biased estimate of the corresponding parameter posterior distribution.

The conditioning threshold u is taken as the 94% quantile of the data, corresponding to a peakflow of $0.2m^3s^{-1}$. The subasymptotic model gives unrealistically small estimates at low levels. The estimator of $\theta(x,12)$ lies far outside of the block bootstrap confidence bounds. This strongly suggests a higher value for u, and this is confirmed by poor estimates of quantiles for the cluster maxima distribution. We thus run the one-step method again, with a threshold at the 98% quantile of the data ($0.4m^3s^{-1}$).

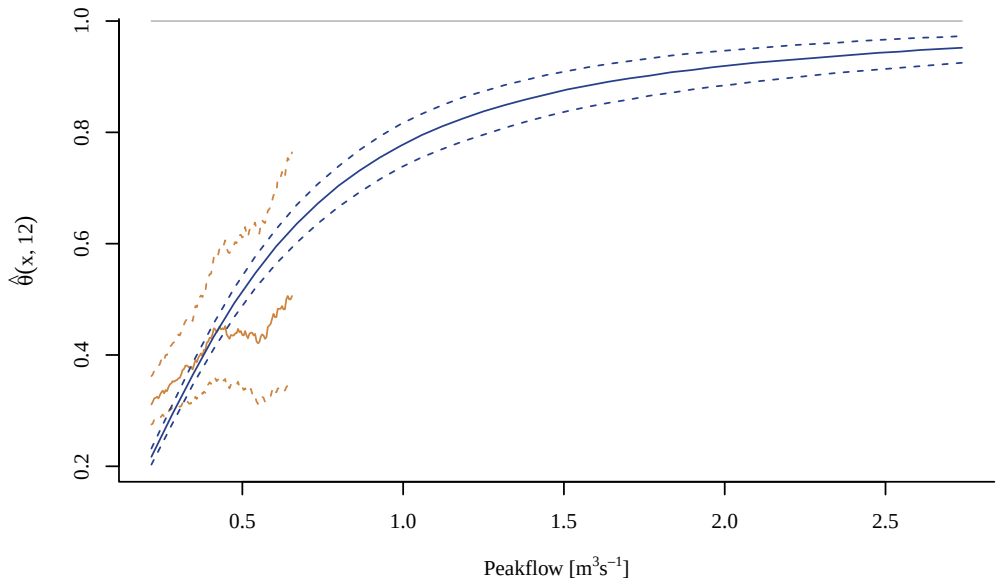

Figure 19 – Runs estimator (solid, light brown) for $\theta(x, 12)$ with block bootstrap confidence intervals (dashed, light brown) on the Lambourn's peakflow data. The corresponding estimator given by the subasymptotic one-step method (solid, blue) and its 2.5% and 97.5% pointwise quantiles (dashed, blue) are superimposed for comparison.

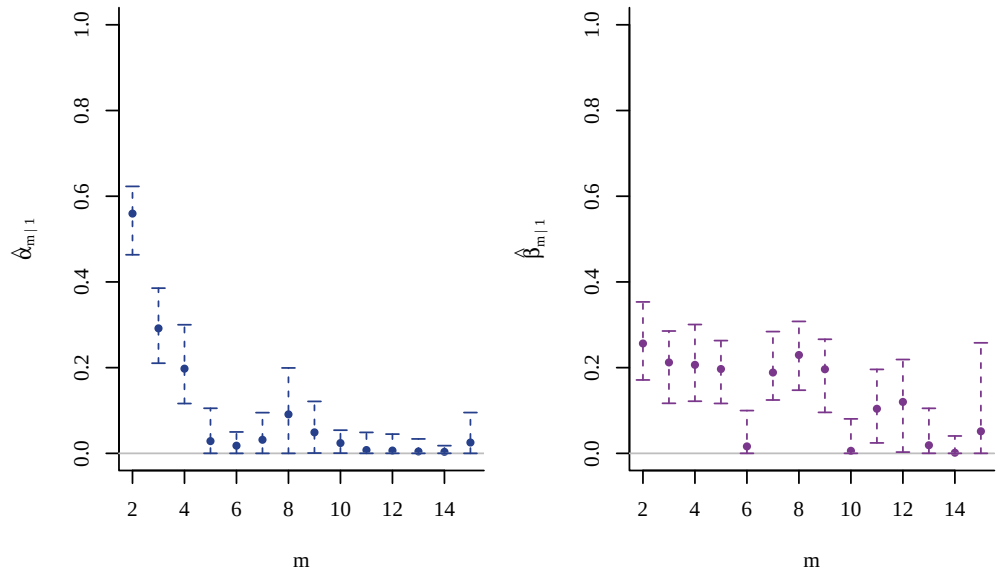

Figure 20 – Median values of the Heffernan–Tawn parameter estimators computed on the last 5,000 iterations of a 15,000 loop run of the one-step method with a conditioning threshold at the 94% quantile. Confidence intervals are also shown and correspond to the 2.5% and 97.5% quantiles of the same output. The horizontal grey lines represent the case of independence, i.e., $\alpha_{m|1} = \beta_{m|1} = 0$.

The stability of $\theta(x,m)$ for large values of x can be studied by considering the ratio $\theta(x,m)/\theta(u,m)$ estimated with the one-step method, so that we directly have confidence intervals to assess if the ratio is significantly larger than 1, for $x \gg u$. Figure 21 illustrates this point for $m = 12$ and values of x ranging from the 98% quantile to the 99.999% quantile; the ratio is significantly far from 1 at all levels.

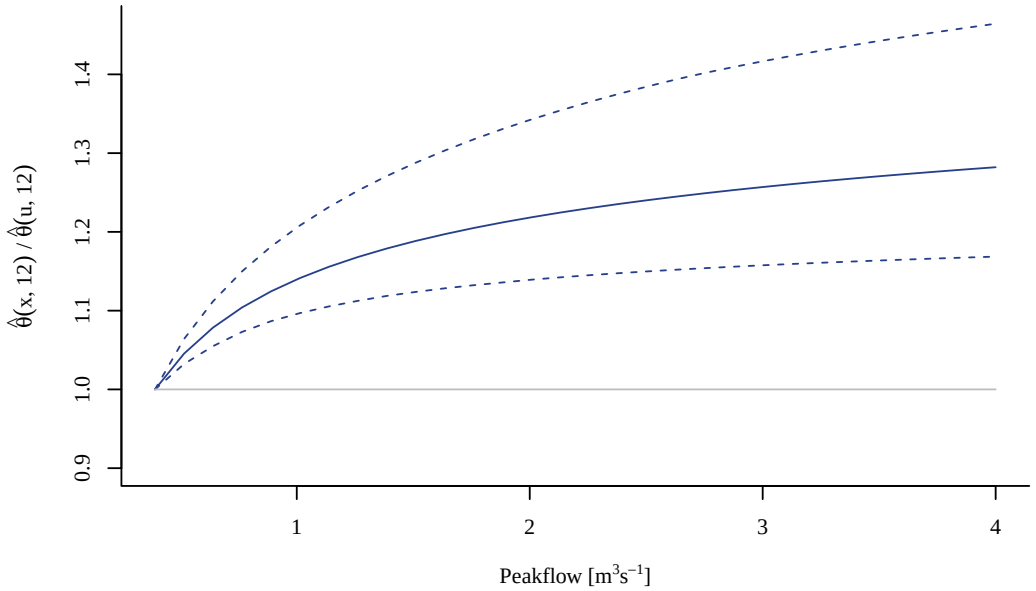

Figure 21 – Estimates of the ratio $\theta(x,12)/\theta(0.4,12)$ with the values of x corresponding to peakflow levels. The dashed lines give a 95% confidence band. The horizontal grey line represents the stability case, i.e., a ratio equal to 1.

To conclude this brief study of the Lambourn's peakflow, we show in Figure 22 the 99% quantile estimated with the peaks over threshold approach and with the subasymptotic model. The confidence interval for the subasymptotic approach has been computed following the procedure detailed in Appendix F. It is wider than the one computed for the peaks over threshold approach, which is based on a delta method. The two estimates only slightly differ compared to the results presented for the simulation study.

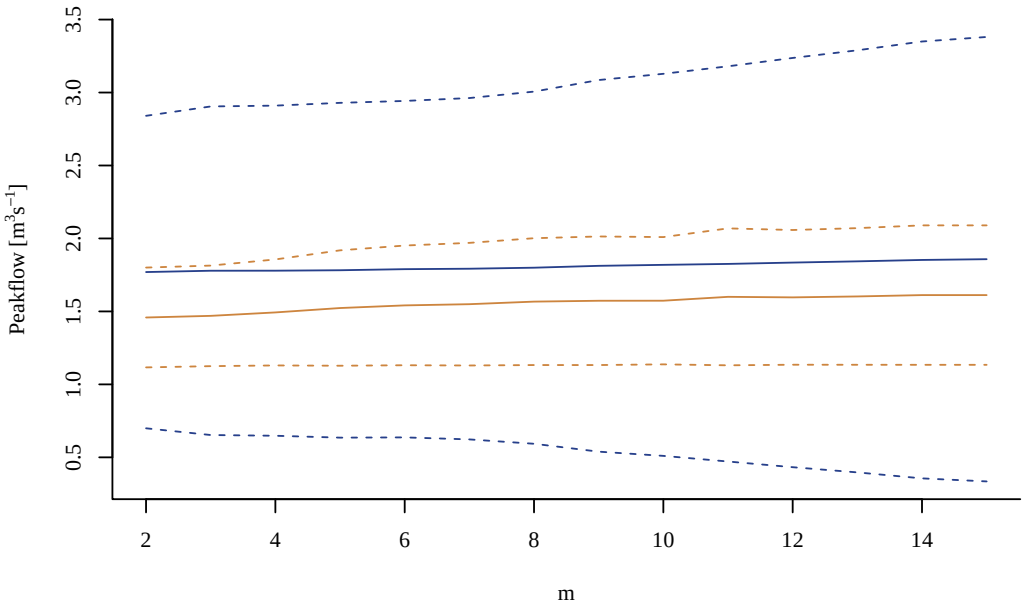

Figure 22 – Quantile at the 99% level estimated with the peaks over threshold approach (light brown) and with the subasymptotic model using the one-step method (solid, blue). The dashed lines give the limits of 95% confidence intervals for both estimates.

Discussion

We have seen how Bayesian semiparametrics can bring much more understanding into a fit of the Heffernan–Tawn model, especially through the fitted residual distribution, which is generally wider than the empirical distribution of the residuals computed from the classical, two-step approach. The confidence intervals for the residual density also translate the uncertainty in the nonparametric part of the model, without any need for bootstrapping the data. This asserts the capacity of the Bayesian approach to capture the uncertainty in the estimates of the parameters $\boldsymbol{\alpha}_{|i}$ and $\boldsymbol{\beta}_{|i}$ in the overall fit.

Another feature of the Dirichlet process is its high flexibility, allowing the parametric and nonparametric parts to compensate each other, the overall fitted density remaining stable through these underlying changes. This can however lead to serious divergences between two- and one-step fits, when comparing their parameters individually. The application of these two methods to the subasymptotic model for cluster maxima involves the whole conditional distribution, leading to very similar results.

Even if the one-step method, as it has been presented, seems to perform well, some improvements and alternatives are discussed below. Some points are further detailed in the appendix.

Sampling $\widehat{\alpha}_{|1}$ and $\widehat{\beta}_{|1}$ As seen in §4.6, the ideal method to sample the Heffernan–Tawn parameters would involve degenerate densities and intractable acceptance ratios. There is no evidence however that the informative beta prior we chose leads to biased estimates. Alternative methods are detailed in appendix C, with a particular attention to the case of $\boldsymbol{\alpha}_{|i}$, for which the posterior density has a closed form.

Within-component dependence As we have seen, the full Bayesian model (4.18) would involve multivariate Gaussian densities to account for dependence within each component of the mixture. The good estimates provided by the simplified approach where each component has independent marginals suggested that the main dependence was already captured through the structure given by the components. Appendix G shows the improvement brought by this simplified approach over the approach where residuals are assumed conditionally independent.

Retrospective sampling The truncation of the stick-breaking representation has been presented as the key point for sampling directly from the posterior of the classification variables c_l, with l ranging through the indices of the observations. A novel technique involving retrospective sampling (Papaspiliopoulos and Roberts, 2008) allows for sampling from the exact, infinite-dimensional stick-breaking process. A direct sampling from the posterior is also presented by the authors. Computationally speaking, we have to account for a potentially very large use of memory and longer running times due to the need for sampling

retrospectively. As the error made by truncating the stick-breaking representation decreases exponentially, and only barely depends on the number of observations, the approximated process seems to be an efficient solution, especially when the Gibbs sampler is part of a bigger model, where speed and memory use become main issues.

Conditional sampling We have mentioned the conditional sampling for the Dirichlet process at the beginning of §4.4 and algorithms using this method. Papaspiliopoulos and Roberts (2008) conducted a systematic study where they compared the efficiency, in terms of *integrated autocorrelation times*, of conditional sampling methods with marginal sampling. The conditional methods have been found to perform better than Gibbs sampling in all cases. This kind of algorithm could be a potentially interesting way of enhancing or at least completing the Gibbs approach.

Covariates We extracted the peakflows from the river flows in the introductory part, and the interest of this work was to focus on the peakflows only, since they reflect the extremal behaviour of river levels. To be complete, the study should consider the entire flow and model the dependence between base- and peakflow; return levels would then correspond to a physical quantity — water flow in $m^3 s^{-1}$. A possible approach is to model the baseflow separately and to set it as a peakflow covariate. Part of this question has been tackled by Jonathan *et al.* (2013), who have generalised the Heffernan–Tawn model to include data about direction of storms in a study about oceanic extreme events.

Acknowledgements

I am especially grateful to Prof. J. A. Tawn for his support and guidance, and for his availability. My thanks also go to Prof. A. C. Davison for his rigorous revisions, and to Prof. P. Fearnhead for his welcome help on label switching.

A Pickands' function: parametric inference

A.1 Asymmetric mixed model

For the asymmetric mixed model, the distribution $F(\boldsymbol{x}) = \exp\{-V(\boldsymbol{x})\}$ has the following expression for V (we assume standard Fréchet margins):

$$V(\boldsymbol{x}) = \frac{x_1 + x_2}{x_1 x_2} - \frac{(\theta+\varphi)/x_1 + (2\varphi+\theta)/x_2}{x_1 x_2 (1/x_1 + 1/x_2)^2}, \quad x_1, x_2 > 0,$$

where $\theta \geq 0$, $\theta + \varphi \leq 1$, $\theta + 2\varphi \leq 1$ and $\theta + 3\varphi \geq 0$. The likelihood contribution for each region as defined in (1.25) and for a bivariate threshold $\boldsymbol{u}$ has general form

$$\begin{aligned}
L_{00}(\boldsymbol{x}) &= \exp\{-V(\boldsymbol{u})\}, && \boldsymbol{x} \in R_{00},\\
L_{10}(\boldsymbol{x}) &= \exp\{-V(x_1,u_2)\}\left\{-\frac{\partial V}{\partial x_1}(x_1,u_2)\right\}, && \boldsymbol{x} \in R_{10},\\
L_{01}(\boldsymbol{x}) &= \exp\{-V(u_1,x_2)\}\left\{-\frac{\partial V}{\partial x_2}(u_1,x_2)\right\}, && \boldsymbol{x} \in R_{01},\\
L_{11}(\boldsymbol{x}) &= \exp\{-V(\boldsymbol{x})\}\left\{\frac{\partial V}{\partial x_1}(\boldsymbol{x})\frac{\partial V}{\partial x_2}(\boldsymbol{x}) - \frac{\partial^2 V}{\partial x_1 \partial x_2}(\boldsymbol{x})\right\}, && \boldsymbol{x} \in R_{11}.
\end{aligned}$$

Partial derivatives of V are

$$\begin{aligned}
\frac{\partial V}{\partial x_1}(\boldsymbol{x}) &= -\frac{(1-\theta-2\varphi)x_1^3 + (3-\theta)x_1^2 x_2 + 3x_1 x_2^2 + x_2^3}{x_1^2 (x_1+x_2)^3},\\
\frac{\partial V}{\partial x_2}(\boldsymbol{x}) &= -\frac{x_1^3 + 3x_1^2 x_2 + (3-\theta-3\varphi)x_1 x_2^2 + (1-\theta-\varphi)x_2^3}{x_2^2 (x_1+x_2)^3},\\
\frac{\partial^2 V}{\partial x_1 \partial x_2}(\boldsymbol{x}) &= -2\frac{\theta(x_1+x_2) + 3\varphi x_1}{(x_1+x_2)^4}.
\end{aligned}$$

The corresponding expression for A is (Tawn, 1988)

$$A(t) = \varphi t^3 + \theta t^2 - (\theta+\varphi)t + 1, \quad 0 \leq t \leq 1. \tag{A.1}$$

A.2 Asymmetric logistic model

This model involves the following formula for V (standard Fréchet margins):

$$V(\boldsymbol{x}) = \frac{1-\theta}{x_1} + \frac{1-\varphi}{x_2} + \left\{\left(\frac{\theta}{x_1}\right)^{1/\alpha} + \left(\frac{\varphi}{x_2}\right)^{1/\alpha}\right\}^{\alpha}, \quad x_1, x_2 > 0,$$

with $0 \leq \theta, \varphi, \alpha \leq 1$.

We give the partial derivatives of V for this model:

$$\begin{aligned}
\frac{\partial V}{\partial x_1}(\boldsymbol{x}) &= -\frac{(\theta/x_1)^{1/\alpha}\left\{(\theta/x_1)^{1/\alpha}+\left(\varphi/x_2\right)^{1/\alpha}\right\}^{\alpha-1}}{x_1}-\frac{1-\theta}{x_1^2},\\
\frac{\partial V}{\partial x_2}(\boldsymbol{x}) &= -\frac{\left(\varphi/x_2\right)^{1/\alpha}\left\{(\theta/x_1)^{1/\alpha}+\left(\varphi/x_2\right)^{1/\alpha}\right\}^{\alpha-1}}{x_2}-\frac{1-\varphi}{x_2^2},\\
\frac{\partial^2 V}{\partial x_1 \partial x_2} &= (\alpha-1)\frac{(\theta/x_1)^{1/\alpha}\left(\varphi/x_2\right)^{1/\alpha}\left\{(\theta/x_1)^{1/\alpha}+\left(\varphi/x_2\right)^{1/\alpha}\right\}^{\alpha-2}}{\alpha x_1 x_2}.
\end{aligned}$$

Finally the related Pickands' function is (Tawn, 1988)

$$A(t) = \left[\{\theta(1-t)\}^{1/\alpha}+\left(\varphi t\right)^{1/\alpha}\right]^{\alpha}+(\theta-\varphi)t+1-\theta, \quad 0 \le t \le 1.$$

B Posterior densities for the multivariate blocked Gibbs sampler

In this appendix and in the following one we slightly simplify the notation and write $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ instead of $\boldsymbol{\mu}_{Z_{|i}}$ and $\boldsymbol{\sigma}^2_{Z_{|i}}$, as well as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ instead of $\boldsymbol{\alpha}_{|i}$ and $\boldsymbol{\beta}_{|i}$. Otherwise the notation remains the same as in the body of the text.

We assume that we have d-dimensional observations $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ such that $X_{i,1}>u$, $i = 1,\ldots,n$, for some threshold $u > 0$. We write $n_k := \sum_{i=1}^n \mathbf{1}(c_i = k)$, the number of observations in component k, and $C_k := \{i \in \{1,\ldots,n\} : c_i = k\}$, the set of indices of observations belonging to component k, so that $n_k = |C_k|$.

B.1 Posterior density for $\boldsymbol{\mu}$

We can compute the posterior density separately for each $\boldsymbol{\mu}_k$, $k = 1,\ldots,N$, as the components are assumed independent. Using Bayes' formula,

$$f\left(\boldsymbol{\mu}_k \mid \boldsymbol{X},\, \boldsymbol{\sigma}_k^2,\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \boldsymbol{\tau}\right) \propto f\left(\boldsymbol{X}_{-1} \mid \boldsymbol{X}_1,\, \boldsymbol{\mu}_k,\, \boldsymbol{\sigma}_k^2,\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \boldsymbol{\tau}\right) f\left(\boldsymbol{\mu}_k \mid \boldsymbol{\tau}\right), \quad k = 1,\ldots,N. \tag{B.1}$$

The prior normal density on the right-hand side of (B.1) is proportional to

$$\exp\left\{-\frac{1}{2}\sum_{j=2}^{d}\frac{\left(\mu_{k,j}-\tau_j\right)^2}{\sigma^2_{\mu,j}}\right\}, \quad k = 1,\ldots,N, \tag{B.2}$$

and the likelihood is of the form

$$\exp\left\{-\frac{1}{2}\sum_{j=2}^{d}\sum_{i\in C_k}\frac{\left(X_{i,j}-\alpha_j X_{i,1}-\mu_{k,j}X_{i,1}^{\beta_j}\right)^2}{X_{i,1}^{2\beta_j}\sigma^2_{k,j}}\right\}, \quad k = 1,\ldots,N. \tag{B.3}$$

The computation of the posterior density involves the product of (B.2) and (B.3), and from

$$\begin{aligned}
&\exp\left[-\frac{1}{2}\sum_{j=2}^{d}\left\{\sum_{i\in C_k}\frac{\left(X_{i,j}-\alpha_j X_{i,1}-\mu_{k,j}X_{i,1}^{\beta_j}\right)^2}{X_{i,1}^{2\beta_j}\sigma^2_{k,j}}+\frac{\left(\mu_{k,j}-\tau_j\right)^2}{\sigma^2_{\mu,j}}\right\}\right] \\
&\quad\propto \exp\left[-\frac{1}{2}\sum_{j=2}^{d}\left\{\sum_{i\in C_k}\frac{\mu^2_{k,j}X_{i,1}^{2\beta_j}-2\mu_{k,j}X_{i,1}^{\beta_j}\left(X_{i,j}-\alpha_j X_{i,1}\right)}{X_{i,1}^{2\beta_j}\sigma^2_{k,j}}+\frac{\mu^2_{k,j}-2\mu_{k,j}\tau_j}{\sigma^2_{\mu,j}}\right\}\right] \\
&\quad= \exp\left[-\frac{1}{2}\sum_{j=2}^{d}\left\{\mu^2_{k,j}\left(\frac{n_k}{\sigma^2_{k,j}}+\frac{1}{\sigma^2_{\mu,j}}\right)-2\mu_{k,j}\left(\frac{1}{\sigma^2_{k,j}}\sum_{i\in C_k}\frac{X_{i,j}-\alpha_j X_{i,1}}{X_{i,1}^{\beta_j}}+\frac{\tau_j}{\sigma^2_{\mu,j}}\right)\right\}\right],
\end{aligned}$$

we deduce, by completing the square, that the posterior density for $\boldsymbol{\mu}_k$ is a multivariate Gaussian density with independent margins, namely

$$\mu_{k,j} \mid \boldsymbol{X}_j,\, \boldsymbol{X}_1,\, \sigma^2_{k,j},\, \alpha_j,\, \beta_j,\, \tau_j \overset{\text{ind}}{\sim} \mathcal{N}\left(M_{\mu_{k,j}}, S^2_{\mu_{k,j}}\right), \quad k = 1,\ldots,N, \quad j = 2,\ldots,d,$$

with parameters

$$M_{\mu_{k,j}} := S^2_{\mu_{k,j}} \left(\frac{1}{\sigma^2_{k,j}} \sum_{i \in C_k} \frac{X_{i,j} - \alpha_j X_{i,1}}{X_{i,1}^{\beta_j}} + \frac{\tau_j}{\sigma^2_{\mu,j}} \right), \quad S^2_{\mu_{k,j}} := \left(\frac{n_k}{\sigma^2_{k,j}} + \frac{1}{\sigma^2_{\mu,j}} \right)^{-1}.$$

B.2 Posterior density for $\boldsymbol{\sigma}^2$

Each $\boldsymbol{\sigma}^2_k$ can be computed separately as the components are independent. We have

$$f\left(\boldsymbol{\sigma}^2_k \mid \boldsymbol{X},\, \boldsymbol{\mu}_k,\, \boldsymbol{\alpha},\, \boldsymbol{\beta}\right) \propto f\left(\boldsymbol{X}_{-1} \mid \boldsymbol{X}_1,\, \boldsymbol{\mu}_k,\, \boldsymbol{\sigma}^2_k,\, \boldsymbol{\alpha},\, \boldsymbol{\beta}\right) f\left(\boldsymbol{\sigma}^2_k\right), \quad k = 1,\ldots,N. \tag{B.4}$$

The inverse-gamma prior density for $\boldsymbol{\sigma}^2_k$ on the right-hand side of (B.4) is, up to a constant,

$$\prod_{j=2}^{d} \sigma_{k,j}^{2(-\nu_{1,j}-1)} \mathrm{e}^{-\nu_{2,j}/\sigma^2_{k,j}}, \quad k = 1,\ldots,N, \tag{B.5}$$

and the likelihood is proportional to

$$\prod_{j=2}^{d} \left[\prod_{i \in C_k} \frac{1}{X_{i,1}^{\beta_j} \sigma_{k,j}} \exp\left\{ -\frac{1}{2} \sum_{i \in C_k} \frac{\left(X_{i,j} - \alpha_j X_{i,1} - \mu_{k,j} X_{i,1}^{\beta_j}\right)^2}{X_{i,1}^{2\beta_j} \sigma^2_{k,j}} \right\} \right], \quad k = 1,\ldots,N, \tag{B.6}$$

so that the posterior density in (B.4) computed through the product of (B.5) and (B.6) leads to

$$\begin{aligned}
&\prod_{j=2}^{d} \left[\prod_{i \in C_k} \frac{\sigma^{-1}_{k,j}}{X_{i,1}^{\beta_j}} \sigma_{k,j}^{2(-\nu_{1,j}-1)} \exp\left\{ -\frac{1}{2} \sum_{i \in C_k} \frac{\left(X_{i,j} - \alpha_j X_{i,1} - \mu_{k,j} X_{i,1}^{\beta_j}\right)^2}{X_{i,1}^{2\beta_j} \sigma^2_{k,j}} - \frac{\nu_{2,j}}{\sigma^2_{k,j}} \right\} \right] \\
&\quad \propto \prod_{j=2}^{d} \left[\sigma_{k,j}^{2(-n_k/2-\nu_{1,j}-1)} \exp\left\{ -\frac{\frac{1}{2} \sum_{i \in C_k} \left(X_{i,j} - \alpha_j X_{i,1} - \mu_{k,j} X_{i,1}^{\beta_j}\right)^2 \Big/ X_{i,1}^{2\beta_j} + \nu_{2,j}}{\sigma^2_{k,j}} \right\} \right],
\end{aligned}$$

from which we conclude that the multivariate posterior density can be split into independent parts, viz.

$$\sigma^2_{k,j} \mid \boldsymbol{X}_j,\, \boldsymbol{X}_1,\, \mu_{k,j},\, \alpha_j,\, \beta_j \overset{\text{ind}}{\sim} \text{Inv-Gamma}\left(N_{1,k,j}, N_{2,k,j}\right), \quad k = 1,\ldots,N, \quad j = 2,\ldots,d,$$

with parameters

$$N_{1,k,j} := \frac{n_k}{2} + \nu_{1,j}, \quad N_{2,k,j} := \frac{1}{2} \sum_{i \in C_k} \frac{\left(X_{i,j} - \alpha_j X_{i,1} - \mu_{k,j} X_{i,1}^{\beta_j}\right)^2}{X_{i,1}^{2\beta_j}} + \nu_{2,j}.$$

B.3 Posterior density for $\boldsymbol{c}$

The general form for the posterior density of the c_i's is given by Bayes' formula:

$$f\left(c_i \mid \boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}\right) \propto f\left(\boldsymbol{X}_{-1} \mid \boldsymbol{X}_1, c_i, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) f(c_i \mid \boldsymbol{w}), \quad i = 1,\ldots,n. \tag{B.7}$$

The stick-breaking prior in (B.7) is

$$\sum_{k=1}^{N} w_k \delta_k(c_i), \quad i = 1,\ldots,n, \tag{B.8}$$

and the related likelihood is proportional to

$$\prod_{j=2}^{d} \left[\frac{1}{X_{i,1}^{\beta_j} \sigma_{c_i,j}} \exp\left\{ -\frac{1}{2} \frac{\left(X_{i,j} - \alpha_j X_{i,1} - \mu_{c_i,j} X_{i,1}^{\beta_j}\right)^2}{X_{i,1}^{2\beta_j} \sigma_{c_i,j}^2} \right\} \right], \quad i = 1,\ldots,n. \tag{B.9}$$

The product of (B.8) with (B.9) gives

$$\begin{aligned} &\prod_{j=2}^{d} \left[\frac{1}{X_{i,1}^{\beta_j} \sigma_{c_i,j}} \exp\left\{ -\frac{1}{2} \frac{\left(X_{i,j} - \alpha_j X_{i,1} - \mu_{c_i,j} X_{i,1}^{\beta_j}\right)^2}{X_{i,1}^{2\beta_j} \sigma_{c_i,j}^2} \right\} \right] \sum_{k=1}^{N} w_k \delta_k(c_i) \\ &\qquad = \sum_{k=1}^{N} \left(\prod_{j=2}^{d} \left[\frac{1}{X_{i,1}^{\beta_j} \sigma_{k,j}} \exp\left\{ -\frac{1}{2} \frac{\left(X_{i,j} - \alpha_j X_{i,1} - \mu_{k,j} X_{i,1}^{\beta_j}\right)^2}{X_{i,1}^{2\beta_j} \sigma_{k,j}^2} \right\} \right] w_k \delta_k(c_i) \right), \end{aligned}$$

which means that the posterior density in (B.7) is such that

$$c_i \mid \boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w} \overset{\text{ind}}{\sim} \sum_{k=1}^{N} W_{k,i} \delta_k, \quad i = 1,\ldots,n,$$

where the stick-breaking weights are defined as

$$W_{k,i} := \frac{w_k}{\overline{W}_i} \prod_{j=2}^{d} \left[\frac{1}{X_{i,1}^{\beta_j} \sigma_{k,j}} \exp\left\{ -\frac{1}{2} \frac{\left(X_{i,j} - \alpha_j X_{i,1} - \mu_{k,j} X_{i,1}^{\beta_j}\right)^2}{X_{i,1}^{2\beta_j} \sigma_{k,j}^2} \right\} \right],$$

with $\overline{W}_i := \sum_{k=1}^{N} W_{k,i}$, $i = 1,\ldots,n$, constants which make the weights sum up to 1.

B.4 Posterior density for $\boldsymbol{w}$

Bayes' formula allows for the following expression related to $\boldsymbol{w}$ posterior density:

$$f\left(\boldsymbol{w} \mid \boldsymbol{c}, \gamma\right) \propto f(\boldsymbol{c} \mid \boldsymbol{w}) f\left(\boldsymbol{w} \mid \boldsymbol{\gamma}\right). \tag{B.10}$$

The prior density in (B.10) is of the form

$$w_N^{\gamma-1} W_2^{-1} \cdots W_{N-1}^{-1}, \tag{B.11}$$

where $W_j := \sum_{k=j}^{N} w_j$, $j = 2,\ldots,N-1$. The other right-hand side density in (B.10) is

$$\prod_{i=1}^{n} \sum_{k=1}^{N} w_k \delta_k(c_i), \tag{B.12}$$

due to conditional independence. The posterior density for $\boldsymbol{w}$ is computed following these steps:

$$\begin{aligned}
&\left\{ \prod_{i=1}^{n} \sum_{k=1}^{N} w_k \delta_k(c_i) \right\} w_N^{\gamma-1} W_2^{-1} \cdots W_{N-1}^{-1} \\
&\quad = \left\{ \prod_{i_1 \in C_1} \sum_{k=1}^{N} w_k \delta_k\left(c_{i_1}\right) \right\} \cdots \left\{ \prod_{i_N \in C_N} \sum_{k=1}^{N} w_k \delta_k\left(c_{i_N}\right) \right\} w_N^{\gamma-1} W_2^{-1} \cdots W_{N-1}^{-1} \\
&\quad = w_1^{n_1} \cdots w_{N-1}^{n_{N-1}} w_N^{n_N+\gamma-1} W_2^{-1} \cdots W_{N-1}^{-1},
\end{aligned}$$

which is proportional to a generalised Dirichlet density of the following form:

$$\boldsymbol{w} \mid \boldsymbol{c}, \gamma \sim \text{GDirichlet}(a_1, b_1, \ldots, a_{N-1}, b_{N-1}),$$
$$a_k := 1 + n_k, \quad b_k := \gamma + \sum_{j=k+1}^{N} n_j, \quad k = 1,\ldots,N-1.$$

B.5 Posterior density for γ

For γ we can write its density as

$$f\left(\gamma \mid \boldsymbol{w}\right) \propto f\left(\boldsymbol{w} \mid \gamma\right) f(\gamma), \tag{B.13}$$

with a gamma prior proportional to

$$\gamma^{\eta_1 - 1} \mathrm{e}^{-\gamma/\eta_2}, \tag{B.14}$$

and a likelihood of the form

$$\left\{ \prod_{k=1}^{N-1} \frac{\Gamma(1+\gamma)}{\Gamma(1)\Gamma(\gamma)} \right\} w_N^{\gamma-1} W_2^{-1} \cdots W_{N-1}^{-1}, \tag{B.15}$$

with $\Gamma(x) := \int_0^\infty u^{x-1} \mathrm{e}^{-u} \,\mathrm{d}u$, $x > 0$. Combining (B.14) and (B.15) we find

$$\begin{aligned}
&\left[\left\{ \prod_{k=1}^{N-1} \frac{\Gamma(1+\gamma)}{\Gamma(1)\Gamma(\gamma)} \right\} w_N^{\gamma-1} W_2^{-1} \cdots W_{N-1}^{-1} \right] \gamma^{\eta_1 - 1} \mathrm{e}^{-\gamma/\eta_2} \\
&\qquad \propto \left(\gamma^{N-1} w_N^{\gamma} \right) \left(\gamma^{\eta-1} \mathrm{e}^{-\gamma/\eta_2} \right) = \gamma^{(N+\eta_1-1)-1} \exp\left(-\gamma \frac{1 - \eta_2 \log w_N}{\eta_2} \right),
\end{aligned}$$

from which we conclude that the posterior density for γ stated in (B.13) is

$$\text{Gamma}\left(N + \eta_1 - 1, \frac{\eta_2}{1 - \eta_2 \log w_N} \right).$$

B.6 Posterior density for $\boldsymbol{\tau}$

Finally, the posterior density for $\boldsymbol{\tau}$ has the form

$$f\left(\boldsymbol{\tau} \mid \boldsymbol{\mu}\right) \propto f\left(\boldsymbol{\mu} \mid \boldsymbol{\tau}\right) f(\boldsymbol{\tau}). \tag{B.16}$$

The normal prior density for $\boldsymbol{\tau}$ is, up to a constant,

$$\exp\left(-\frac{1}{2}\sum_{j=2}^{d}\frac{\tau_j^2}{\sigma_{\tau_j}^2}\right), \tag{B.17}$$

and the likelihood in (B.16) is proportional to

$$\exp\left\{-\frac{1}{2}\sum_{j=2}^{d}\sum_{k=1}^{N}\frac{\left(\mu_{k,j}-\tau_j\right)^2}{\sigma_{\mu_j}^2}\right\}. \tag{B.18}$$

The posterior is simply derived from the product of the two normal densities in (B.17) and (B.18):

$$\begin{aligned}
&\exp\left[-\frac{1}{2}\sum_{j=2}^{d}\sum_{k=1}^{N}\left\{\frac{\left(\mu_{k,j}-\tau_j\right)^2}{\sigma_{\mu_j}^2}\right\}+\frac{\tau_j^2}{\sigma_{\tau_j}^2}\right] \\
&\qquad\propto \prod_{j=2}^{d}\exp\left\{-\frac{1}{2}\left(\sum_{k=1}^{N}\frac{\tau_j^2-2\tau_j\mu_{k,j}}{\sigma_{\mu_j}^2}+\frac{\tau_j^2}{\sigma_{\tau_j}^2}\right)\right\} \\
&\qquad= \prod_{j=2}^{d}\exp\left[-\frac{1}{2}\left\{\tau_j^2\left(\frac{N}{\sigma_{\mu_j}^2}+\frac{1}{\sigma_{\tau_j}^2}\right)-2\tau_j\sum_{k=1}^{N}\frac{\mu_{k,j}}{\sigma_{\mu_j}^2}\right\}\right].
\end{aligned}$$

By completing the square in the exponent, we get a multivariate normal posterior density with independent margins, so that

$$\tau_j \mid \boldsymbol{\mu}_j \overset{\text{ind}}{\sim} \mathcal{N}\left\{\left(\frac{N}{\sigma_{\mu_j}^2}+\frac{1}{\sigma_{\tau_j}^2}\right)^{-1}\sum_{k=1}^{N}\frac{\mu_{k,j}}{\sigma_{\mu_j}^2}, \left(\frac{N}{\sigma_{\mu_j}^2}+\frac{1}{\sigma_{\tau_j}^2}\right)^{-1}\right\}, \quad j=2,\ldots,d.$$

C Alternative sampling methods

As seen in §4.6, sampling from posterior distributions with bounded support, especially when the posterior distribution is degenerate at some points, cannot be implemented directly with a Metropolis–Hastings algorithm. We present a method which translates the question to a posterior with infinite support. In the second part of this appendix, alternative methods are presented to directly sample from the posterior distribution of $\boldsymbol{\alpha}$.

C.1 Transformation to unconstrained posterior

Expanding the unit interval to the real line can be achieved with a simple *logit* transformation:

$$\begin{aligned} \psi : \; [0,1] &\longrightarrow \mathbb{R} \\ \theta &\longmapsto \log\left(\frac{\theta}{1-\theta}\right) =: \tau, \end{aligned}$$

with inverse mapping

$$\begin{aligned} \psi^{-1} : \; \mathbb{R} &\longrightarrow [0,1] \\ \tau &\longmapsto \frac{\mathrm{e}^{\tau}}{1+\mathrm{e}^{\tau}} =: \theta. \end{aligned}$$

The idea brought by this transformation is to use a Metropolis–Hastings step to sample a value for $\tau \in \mathbb{R}$ instead of $\theta \in [0,1]$, and to then transform the output back onto the unit interval using ψ^{-1}. The following relation holds:

$$f_{\tau}(x) = f_{\theta}\left\{\psi^{-1}(x)\right\} \left|\frac{\mathrm{d}\psi^{-1}(x)}{\mathrm{d}x}\right|, \quad x \in \mathbb{R},$$

with Jacobian

$$\left|\frac{\mathrm{d}\psi^{-1}(x)}{\mathrm{d}x}\right| = \frac{\mathrm{e}^{x}}{(1+\mathrm{e}^{x})^{2}}, \quad x \in \mathbb{R}.$$

The acceptance ratio at time t for θ, that is,

$$\frac{\ell\left(\boldsymbol{X} \mid \theta^{(\star)}\right) f_{\theta}\left(\theta^{(\star)}\right)}{\ell\left(\boldsymbol{X} \mid \theta^{(t-1)}\right) f_{\theta}\left(\theta^{(t-1)}\right)},$$

can then be stated in terms of τ as

$$\frac{\ell\left\{\boldsymbol{X} \mid \psi^{-1}\left(\tau^{(\star)}\right)\right\} f_{\tau}\left(\tau^{(\star)}\right)}{\ell\left\{\boldsymbol{X} \mid \psi^{-1}\left(\tau^{(t-1)}\right)\right\} f_{\tau}\left(\tau^{(t-1)}\right)} \frac{\mathrm{e}^{\tau^{(\star)}}\left(1+\mathrm{e}^{\tau^{(t-1)}}\right)^{2}}{\mathrm{e}^{\tau^{(t-1)}}\left(1+\mathrm{e}^{\tau^{(\star)}}\right)^{2}},$$

with $\boldsymbol{X}$ representing the data and ℓ the log-likelihood function. We assume a symmetric proposal density in both cases, such that it does not appear in this explanation.

The issue of boundary masses could be circumvent by computer rounding when transforming τ back to θ, or by manually imposing limits on the real line, beyond which every value of τ is set to $\pm\infty$.

C.2 Direct sampling

While finding an analytical form for the $\boldsymbol{\beta}$ posterior distribution is hopeless, the case of $\boldsymbol{\alpha}$ leads to a non-conjugate posterior distribution which opens the door to a special case of direct sampling, discussed afterwards.

C.2.1 Posterior density for $\boldsymbol{\alpha}$

We can derive the type of the posterior density for $\boldsymbol{\alpha}$, given

$$f\left(\boldsymbol{\alpha} \mid \boldsymbol{X},\, \boldsymbol{\mu},\, \boldsymbol{\sigma}^2,\, \boldsymbol{\beta},\, \boldsymbol{c}\right) \propto f\left(\boldsymbol{X}_{-1} \mid \boldsymbol{X}_1,\, \boldsymbol{\mu},\, \boldsymbol{\sigma}^2,\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \boldsymbol{c}\right) f(\boldsymbol{\alpha}). \tag{C.1}$$

With a uniform prior density equal to

$$\mathbf{1}\left\{\boldsymbol{\alpha} \in [0,1]^{d-1}\right\}, \tag{C.2}$$

and a likelihood proportional to

$$\prod_{j=2}^{d} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \frac{\left(X_{i,j} - \alpha_j X_{i,1} - \mu_{c_i,j} X_{i,1}^{\beta_j}\right)^2}{X_{i,1}^{2\beta_j}\sigma_{c_i,j}^2}\right\}. \tag{C.3}$$

With both (C.2) and (C.3) we can find the form of the posterior for $\boldsymbol{\alpha}$ in (C.1):

$$\begin{aligned}
&\prod_{j=2}^{d} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \frac{\left(X_{i,j} - \alpha_j X_{i,1} - \mu_{c_i,j} X_{i,1}^{\beta_j}\right)^2}{X_{i,1}^{2\beta_j}\sigma_{c_i,j}^2}\right\} \mathbf{1}\left\{\boldsymbol{\alpha} \in [0,1]^{d-1}\right\} \\
&\quad = \prod_{j=2}^{d} \left[\exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \frac{\left(X_{i,j} - \alpha_j X_{i,1} - \mu_{c_i,j} X_{i,1}^{\beta_j}\right)^2}{X_{i,1}^{2\beta_j}\sigma_{c_i,j}^2}\right\} \mathbf{1}\left\{\alpha_j \in [0,1]\right\}\right] \\
&\quad \propto \prod_{j=2}^{d} \left[\exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \frac{\alpha_j^2 X_{i,1}^2 - 2\alpha_j X_{i,1}\left(X_{i,j} - \mu_{c_i,j} X_{i,1}^{\beta_j}\right)}{X_{i,1}^{2\beta_j}\sigma_{c_i,j}^2}\right\} \mathbf{1}\left\{\alpha_j \in [0,1]\right\}\right] \\
&\quad = \prod_{j=2}^{d} \left[\exp\left\{-\frac{1}{2}\left(\alpha_j^2 \sum_{i=1}^{n} \frac{X_{i,1}^2}{X_{i,1}^{2\beta_j}\sigma_{c_i,j}^2} - 2\alpha_j \sum_{i=1}^{n} X_{i,1} \frac{X_{i,j} - \mu_{c_i,j} X_{i,1}^{\beta_j}}{X_{i,1}^{2\beta_j}\sigma_{c_i,j}^2}\right)\right\} \mathbf{1}\left\{\alpha_j \in [0,1]\right\}\right].
\end{aligned}$$

We complete the square and find that an independent posterior marginal for α_j is a truncated normal with mean M_{α_j} and variance $S_{\alpha_j}^2$, with these parameters being defined as

$$M_{\alpha_j} := S_{\alpha_j}^2 \left(\sum_{i=1}^{n} \frac{X_{i,j} - \mu_{c_i,j} X_{i,1}^{\beta_j}}{X_{i,1}^{2\beta_j - 1}\sigma_{c_i,j}^2}\right), \quad S_{\alpha_j}^2 := \left(\sum_{i=1}^{n} \frac{1}{X_{i,1}^{2(\beta_j - 1)}\sigma_{c_i,j}^2}\right)^{-1}.$$

C.2.2 Sampling from a truncated distribution

The basic method for sampling from a truncated distribution can be described in three steps. First compute the probabilities at the boundaries of the truncation interval, then generate a value from a uniform with these probabilities as its support, finally map the generated value to the original scale by applying the — generalised — inverse of the distribution. Mathematically speaking, if the distribution from which we want to generate a sample is $\mathbf{1}\{q_0 \leq \cdot \leq q_1\}F$ the steps described above correspond to:

- set $p_0 := F(q_0)$ and $p_1 := F(q_1)$,
- generate $U \sim \mathscr{U}(p_0, p_1)$,
- set $X := F^{-1}(U)$,

then X is distributed as $\mathbf{1}\{q_0 \leq \cdot \leq q_1\}F$. This procedure is depicted in Figure 23 in the case of a truncation close to the centre of the distribution F. It fails however when $p_1 - p_0$ becomes very small, which typically happens for $q_0 < q_1 \ll \mu_F$ or $\mu_F \ll q_0 < q_1$, with μ_F the mean of F, i.e., when the truncation is in a tail of F. In such a case, computer approximations lead to $p_1 = p_0 \in \{0, 1\}$, and the output X ends up to be the lower, respectively the upper endpoint of F, which does generally not belong to $[q_0, q_1]$.

This is an issue encountered when trying to sample from the posterior distribution of $\boldsymbol{\alpha}$. The next two sections deal with methods which try to overcome this and are more efficient in the case of tail truncation.

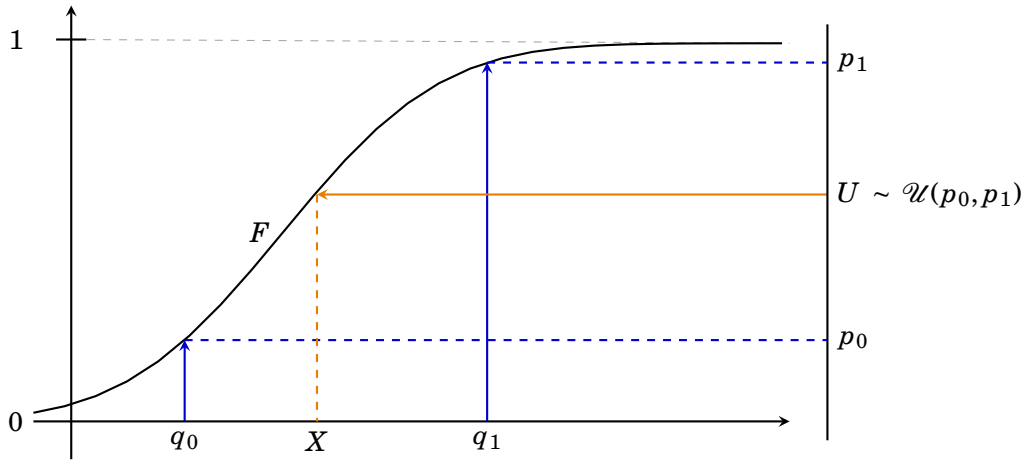

Figure 23 – Illustration of the basic principle for sampling from a distribution F truncated on the interval $[q_0, q_1]$.

C.2.3 Mills' ratio

The Mill's ratio gives an approximation to the normal distribution in terms of its density which may be exploited to compute tail probabilities. Using the normal density instead of the normal distribution allows to manipulate a closed form formula, and to gain in accuracy. By integrating by parts the normal integral after having amplified its integrand appropriately, we get the relation

$$1-\Phi(x) \;=\; \frac{\varphi(x)}{x} + O\left\{\frac{\varphi(x)}{x^{-3}}\right\}, \tag{C.4}$$

where the last term tends quickly to 0 for x sufficiently large, leading to an accurate approximation of the left-hand tail. Using the symmetry of the normal distribution, we can always modify our problem in order to sample from the left-hand tail of a truncated standard normal distribution, by applying

$$\text{sign}(\mu - q_0)\frac{q_i - \mu}{\sigma}, \quad i \;=\; 1,2,$$

where we assume $\text{sign}(\mu - q_0) \;=\; \text{sign}(\mu - q_1)$, otherwise a basic sampling method can be applied. We can then transform the sampled value X back to the original scale and original tail through

$$\text{sign}(\mu - q_0)X\sigma + \mu. \tag{C.5}$$

We now consider a standard normal truncated in its left-hand tail, having support $[q_0, q_1]$. Following the procedure presented in §C.2.2, we can compute p_0 and p_1 using the approximation $\Phi(q_i) \;\approx\; \varphi(q_i)/q_i$ derived from (C.4). The second step consists in generating U from a uniform distribution $\mathscr{U}(p_0, p_1)$. For the last step, we have to find the inverse transformation from the uniform to the Gaussian scale. We have to find X such that $U \;=\; \varphi(X)/X$. Taking the logarithm on both sides, we get a first approximation:

$$\log(U) \;=\; -\frac{X^2}{2} - \log(X) - \frac{1}{2}\log(2\pi) \tag{C.6}$$

$$\Rightarrow \qquad X \;=\; \sqrt{-2\log(U)} + \varepsilon, \tag{C.7}$$

with $\varepsilon \;:=\; \varepsilon(X)$ growing slowly in X compared to X. We inject (C.7) into (C.6) to get

$$\begin{aligned}\log(U) \;&=\; -\frac{1}{2}\left[-2\log(U) + 2\varepsilon\sqrt{-2\log(U)} + \varepsilon^2\right] - \log\left\{\sqrt{-2\log(U)} + \varepsilon\right\} - \frac{1}{2}\log(2\pi) \\ &\approx\; \log(U) - \varepsilon\sqrt{-2\log(U)} - \frac{1}{2}\log\{-2\log(U)\} - \frac{1}{2}\log(2\pi),\end{aligned}$$

so that we have the following expression for ε:

$$\varepsilon \;\approx\; -\frac{\log\{-2\log(U)\} + \log(2\pi)}{2\sqrt{-2\log(U)}}. \tag{C.8}$$

Combining (C.7) with (C.8), we can compute X as follows:

$$X \;=\; -\frac{4\log(U) + \log\{-2\log(U)\} + \log(2\pi)}{2\sqrt{-2\log(U)}}.$$

C.2.4 Rejection method

The rejection method is derived from a fundamental property for densities (Devroye, 1986, §II.3) which suggests the following procedure: assume we wish to generate X from F, with density f satisfying

$$f(x) \leq cg(x), \quad x \in \text{supp}(f),\ c > 0, \tag{C.9}$$

where $\text{supp}(f) \subseteq \text{supp}(g)$. Then X in the following *rejection algorithm* is distributed according to F:

(i) generate Y having density g,

(ii) independently generate U from a uniform distribution,

(iii) set $X = Y$ if $U \leq \dfrac{f(Y)}{cg(Y)}$; otherwise go to (i).

Write N the number of loops needed to generate X, and p the probability of acceptance in step (iii). The constant c has to be the smallest one satisfying (C.9), since N is geometric with probability of success p, and

$$\begin{aligned} p &= \Pr\{Ucg(Y) \leq f(Y)\} \\ &= \int_{\text{supp}(g)} \Pr\{Ucg(Y) \leq f(Y) \mid Y = y\} \Pr(Y = y)\,\mathrm{d}y \\ &= \int_{\text{supp}(g)} \Pr\left\{U \leq \frac{f(y)}{cg(y)}\right\} g(y)\,\mathrm{d}y \\ &= \frac{1}{c}\int_{\text{supp}(f)} f(y)\,\mathrm{d}y = \frac{1}{c}, \end{aligned}$$

so that $\mathrm{E}(N) = 1/p = c$ and $\text{var}(N) = (1-p)/p^2 = c(c-1)$ (Devroye, 1986).

As in the previous section, we can assume without loss of generality that we wish to generate samples from a standard normal distribution with truncation in its left-hand tail. To overcome the issue of having small probability of the standard normal on $[q_0, q_1]$, Robert (1995) proposed a version of the rejection algorithm where the *envelope density* g is chosen as uniform on $[q_0, q_1]$. But the acceptance probability is still very small when q_1 is far below -2 and $q_1 - q_0$ is close to 0. We propose an alternative method using a shifted negative exponential $\text{Exp}(1/2, q_1)$, whose corresponding density function is $\lambda \exp\{\lambda(x - q_1)\}$.

We first find c by bounding the ratio of densities

$$\frac{f(x)}{g(x)} = \frac{1}{\lambda\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2} - \lambda(x - q_1)\right\} \leq \frac{1}{\lambda\sqrt{2\pi}} \exp\left(\frac{\lambda^2}{2} + \lambda q_1\right) =: c, \quad x \in [q_0, q_1].$$

We now have to find the value of λ that minimises c, and it boils down to finding the root of

$$\lambda + q_1 - \frac{1}{\lambda} = 0, \quad \lambda > 0,$$

which is $\lambda = \left(-q_1 + \sqrt{q_1^2+4}\right)/2$. The ratio in step (iii) of the rejection algorithm can be simplified by using the expression we found for c:

$$\frac{f(x)}{cg(x)} = \exp\left\{-\frac{x^2}{2} - \lambda(x-q_1) - \frac{\lambda^2}{2} - \lambda q_1\right\} = \exp\left\{-\frac{1}{2}(x+\lambda)^2\right\}, \quad x \in [q_0, q_1].$$

By choosing a variant of a truncated exponential density as the envelope function g, the corresponding sampling in step (i) can be done using the basic method seen in §C.2.2. There is no need for an approximation since the distribution function has a closed form: we can directly compute $p_i = \lambda e^{\lambda(q_i - q_1)}$, $i = 1,2$, using the optimal λ as discussed above.

D Gibbs sampler output

This section is dedicated to plots of the output of the Gibbs sampler in §5.2. The algorithm has been running through 20,000 loops following a 5,000 iteration burn-in, with a maximum number of components $N = 150$, on five peakflow series conditioned on the 3% largest observations of the Thames' peakflow.

A first comment can be made on the efficiency of label switching in this context where the underlying data are not well-separated into components: the similar sizes of the main components make the ordering hard to find for the Gibbs sampler. Figure 24 summarises the situation, with the two main components having very similar sizes, hence the significant difference between the density of the maximum weight and the density of the first component.

In Figure 25, examples of traces for $\boldsymbol{\alpha}_{|1}$ and $\boldsymbol{\beta}_{|1}$ illustrate the phenomenon of compensation between the two parameters and the residual density. We can identify two different types of behaviour, the first one being a sudden and short transfer, here between $\widehat{\beta}_{2|1}$ and $\widehat{\boldsymbol{\mu}}_{Z_{|1}}$, not shown here, around the 10,000th iteration, followed by a decrease in both $\widehat{\alpha}_{2|1}$ and $\widehat{\beta}_{2|1}$, when the residual mean of the main components starts increasing. The second type of transfer is on a longer time period, and seems to involve only the parametric part of the model, e.g. between $\widehat{\alpha}_{3|1}$ and $\widehat{\beta}_{3|1}$, with opposite trends beginning from the 15,000th iteration onwards.

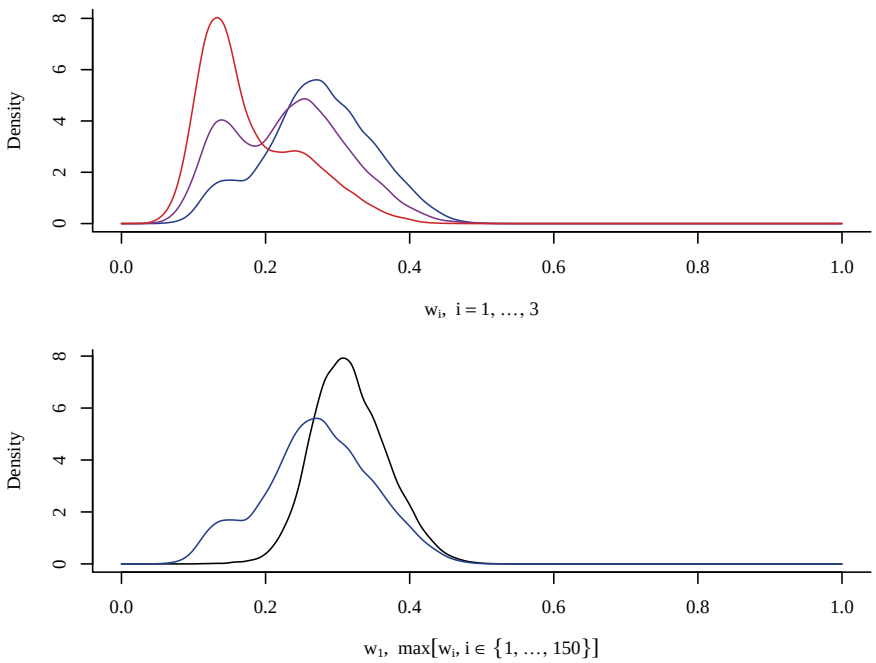

Figure 24 – Top panel: prior weight densities of the first 3 components computed on the 20,000 iterations of the Gibbs sampler output. Bottom panel: weight density of the first component compared with the density of the maximum weight.

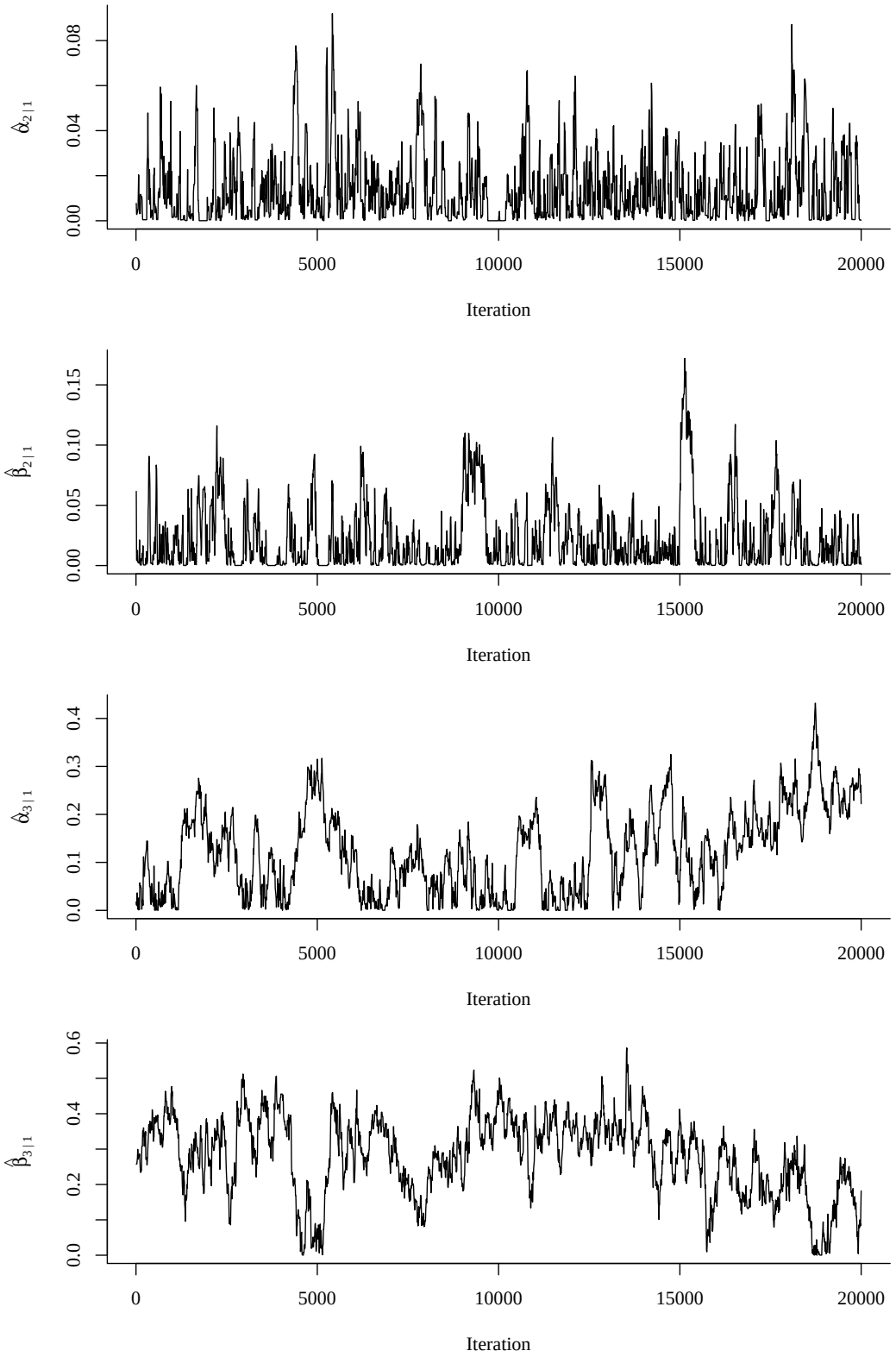

Figure 25 – Traces for $\widehat{\alpha}_{j|1}$ and $\widehat{\beta}_{j|1}$, $j = 2,3$, corresponding to the river peakflows from the Ray and the Lambourn conditioned on the Thames' peakflow.

Figure 26 gives the four residual marginal densities which have not been shown in the main text. It is a comparison between the mean of normal mixtures of the Bayesian approach and a kernel smoothing of the residuals from the frequentist method. The density estimated through the two-step method is systematically more squeezed than the one estimated with the one-step method, due to underestimating the first-step uncertainty.

A couple of bivariate joint densities conditioned on high values of X_1 are shown in Figure 27. Joint density estimated from the fitted Heffernan–Tawn model are compared to the density of the observations. Kernel density estimation is applied to data simulated from the fitted model. We can observe the variation of the contour lines across iterations, each based on samples of size $1,500$. The fitted density is compared with the kernel smoothing approximation of the corresponding peakflows.

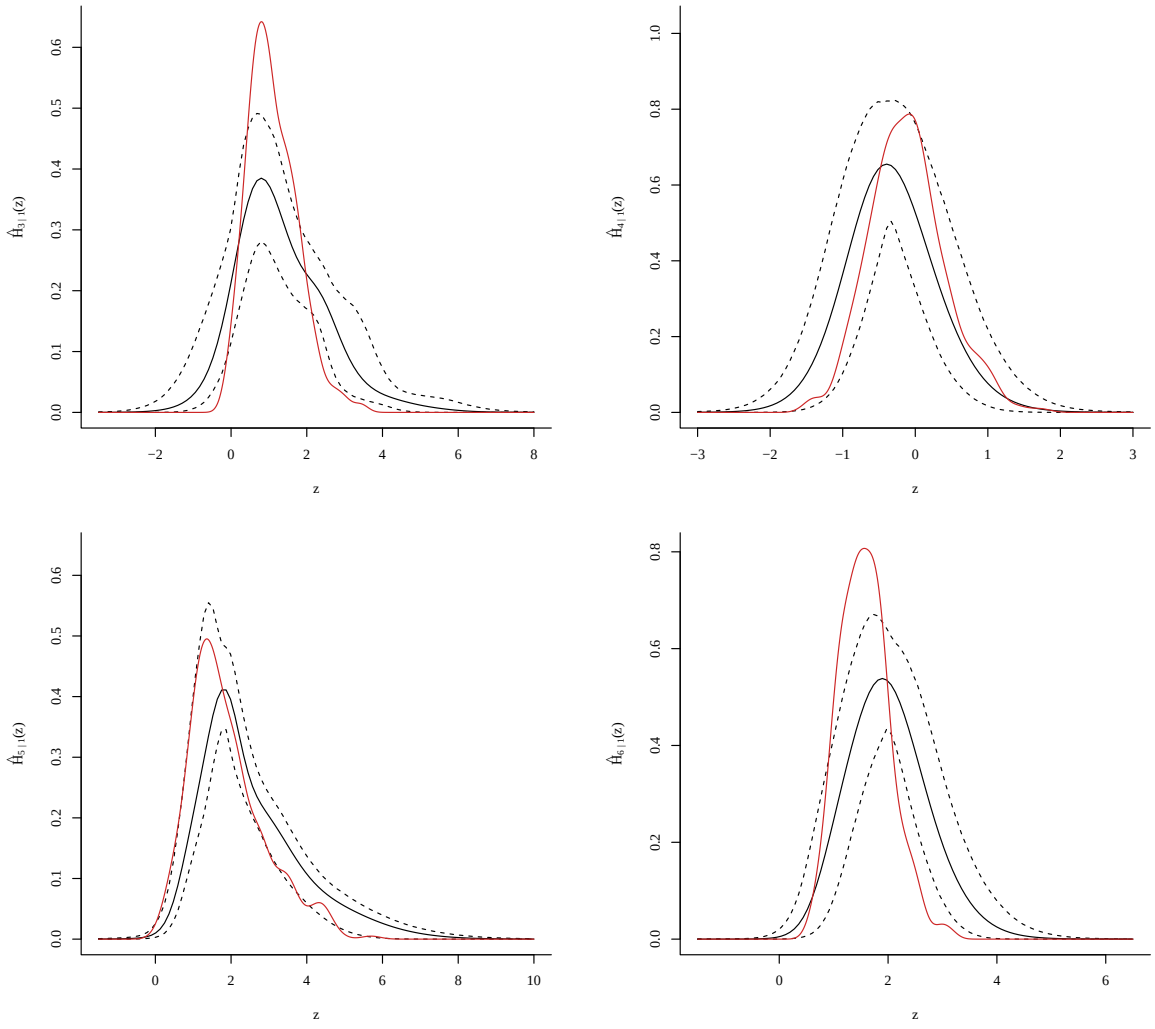

Figure 26 – From top down and from left to right: the marginal residual densities of the Lambourn's, the Coln's, the Mole's and the Ock's peakflows conditioned on high levels of the Thames' peakflow. The black, solid lines are the pointwise means computed on 2,000 iterations randomly selected from the Gibbs sampler output, with their corresponding 2.5% and 97.5% pointwise quantiles (dashed). The red curves are kernel densities adjusted on the residuals computed in the second step of the likelihood approach.

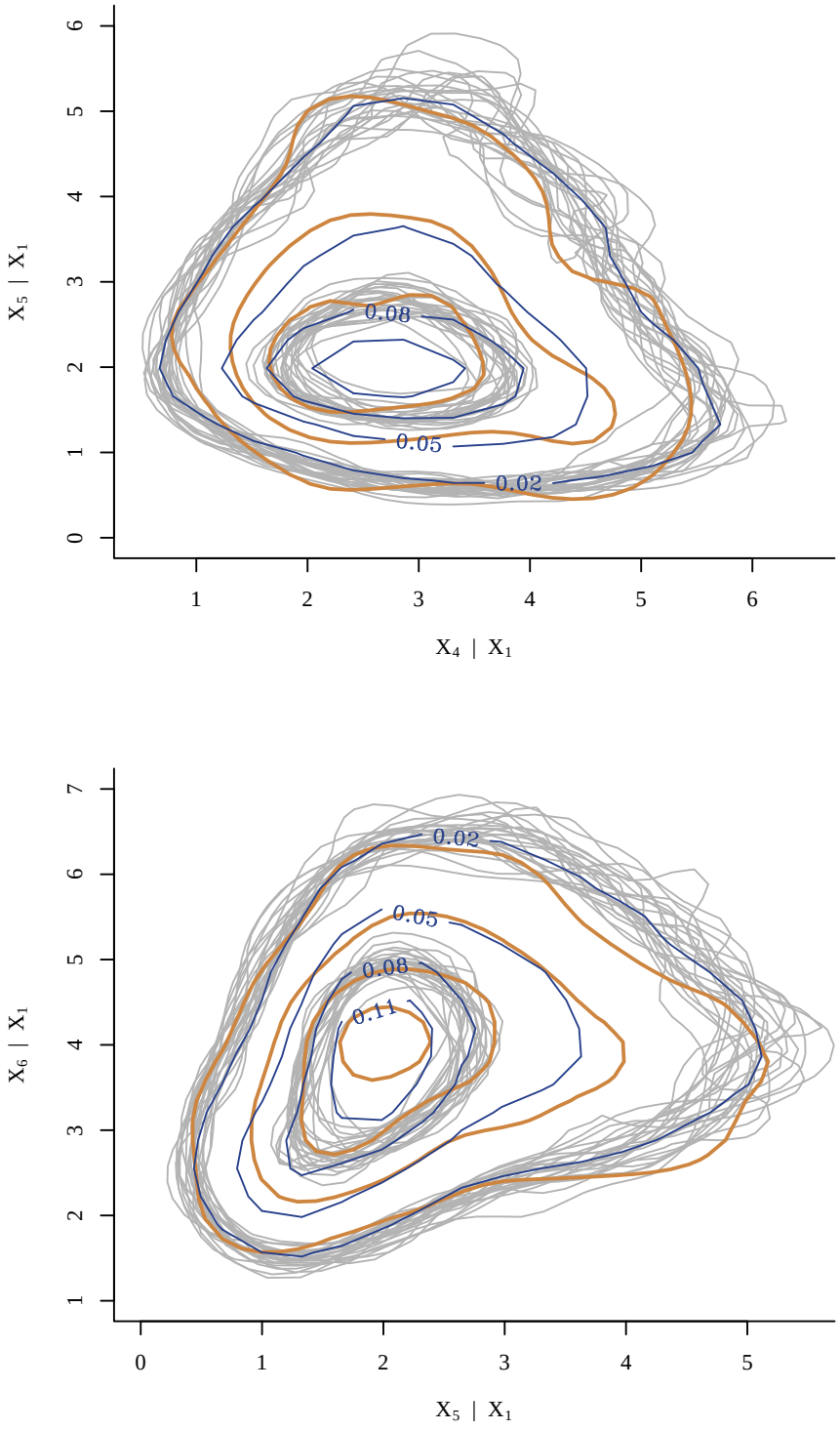

Figure 27 – Contours of bivariate densities conditioned on high levels of X_1. Grey lines are contour lines for 25 randomly chosen iterations from the Gibbs sampler output, corresponding to densities of 0.02 and 0.08. Each one is based on 1,500 simulated data. Blue contours come from the kernel density estimated on 300 such iterations. The related contours computed from the original data are shown in light brown.

E High-dimensional calculus: improving computational efficiency

As seen previously, estimating $\theta(x,m)$ in the one-step framework leads to multiple nested loops over iterations, sample size and dimensions. In this appendix we present a way to make use of tensors and matrices in order to reduce dramatically these computationally intensive control structures.

A tensor is defined as a multidimensional array. A first-order tensor is a vector, a second-order tensor a matrix and for clarity we simply call *tensor* a third-order tensor. In this appendix we will mostly use the notation as proposed by Kiers (2000), i.e., vectors are written as lowercase boldface letters, e.g. $\boldsymbol{v}$, matrices as uppercase boldface letters, e.g. $\boldsymbol{M}$, and tensors as uppercase curly letters, e.g. $\mathscr{T}$. An element $(i_1,\ldots,i_n)$ of an nth-order tensor $\mathscr{T}$ is written as the corresponding lowercase letter, e.g. $t_{i_1,\ldots,i_n}$. In this context we speak about *modes* instead of *dimensions* (Tucker, 1966), and refer to mode A, B and C for the columns, rows and tubes. To refer to a subset of modes of a tensor, we denote by a semicolon the elements which are unfixed, i.e., the front matrix of a tensor $\mathscr{T}$ is $\boldsymbol{T}_{::1}$. See Figure 28 for a graphical representation.

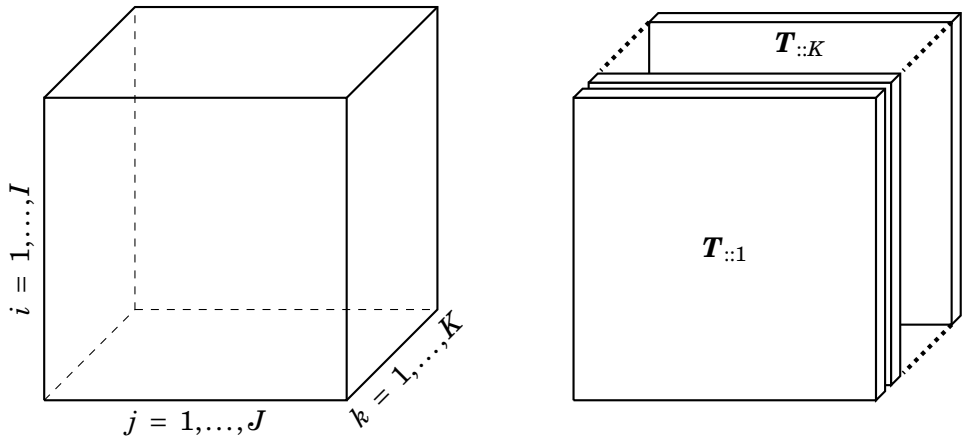

Figure 28 – Left panel: representation of a tensor $\mathscr{T} \in \mathbb{R}^{I\times J\times K}$. Right panel: the same tensor split into slices $\boldsymbol{T}_{::k}$, $k = 1,\ldots,K$, where some are omitted for clarity.

We now define two important operations which allow for flattening a tensor into lower modes. *Vectorisation* is the transformation of a matrix into a vector. The vector ordering is not relevant, provided that it is consistent throughout calculations (Kolda and Bader, 2009). We fix it to be the concatenation of the matrix columns. More precisely if $\boldsymbol{M} \in \mathbb{R}^{I\times J}$ its vectorisation $\boldsymbol{v} \in \mathbb{R}^{IJ}$ is

$$\operatorname{vec}(\boldsymbol{M})_{i+(j-1)I} := v_{i+(j-1)I} = m_{i,j}, \quad i = 1,\ldots,I, \quad j = 1,\ldots,J. \tag{E.1}$$

Figure 29 shows the vectorisation as we define it here.

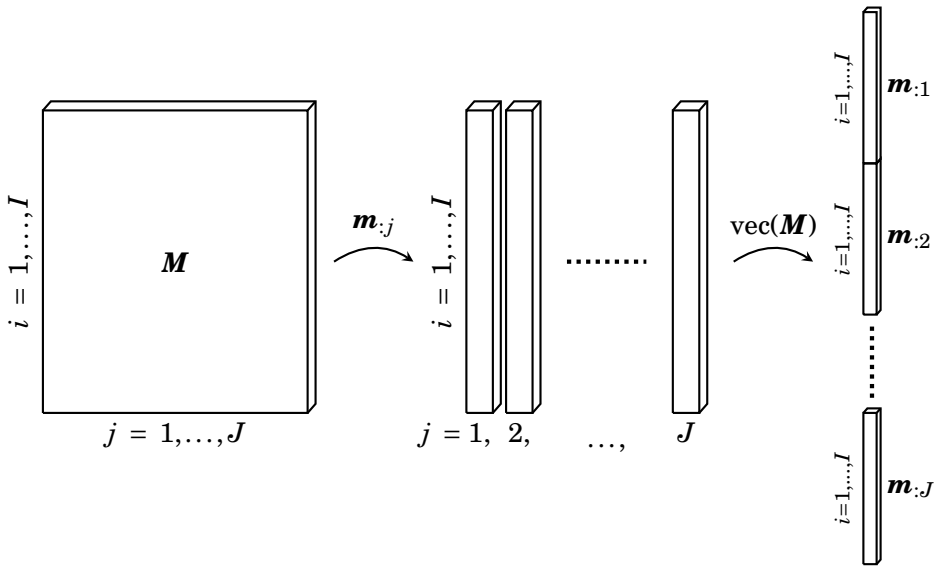

Figure 29 – Vectorisation of an $(I \times J)$-matrix into a vector of length IJ.

The second operation is the flattening of tensors into matrices, called *matricisation*. In the same way as for the vectorisation transformation, we define the matricisation in a restrictive way, namely the mode-1 matricisation, and with an ordering consistent with the vectorisation (E.1). The matricisation $\boldsymbol{T}_{(1)} \in \mathbb{R}^{I \times JK}$ of a tensor $\mathcal{T} \in \mathbb{R}^{I \times J \times K}$ can be defined through the mapping between element (i,j,k) of $\mathcal{T}$ and element (i,l) of $\boldsymbol{T}_{(1)}$, with $l = j + (k-1)J$. Figure 30 is a graphical representation of this transformation.

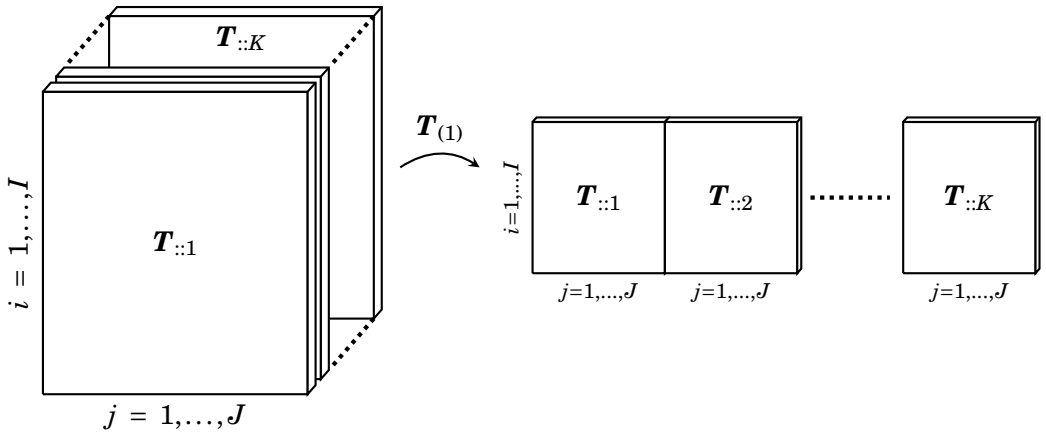

Figure 30 – Matricisation of an $(I \times J \times K)$-tensor into a matrix with modes $I \times JK$.

Coming back to our algorithm used to estimate $\theta(x,m)$ in § 6.1.2, the most expensive part in terms of computational complexity is the computation of the conditioned variable written as $\boldsymbol{X}_{G,2:d}$ in the body of the text and which we write $\boldsymbol{x}_{-1}$ here in

order to comply with the notation introduced above. The sample of values for $\boldsymbol{x}_{-1}$ we want to generate is three-modal, with modes $R \times (d-1) \times S$, where R is the original sample size of the conditioning variable and S is the number of iterations considered. We have at our disposal — according to the notation introduced in this appendix — a tensor $\mathcal{Z}_{|1} \in \mathbb{R}^{R \times (d-1) \times S}$ of sampled residuals, a vector $\boldsymbol{x}_1 \in \mathbb{R}^R$ of sampled conditioning variables and two matrices of parameters $\boldsymbol{A}$, $\boldsymbol{B} \in \mathbb{R}^{(d-1) \times S}$.

The main idea is to matricise the tensors and vectorise the matrices, and to do the computations within a lower-modal context. We finally put the result back into tensor form for further manipulations. Computing the matricised version of $\mathcal{X}_{-1}$ can be done in one line:

$$\boldsymbol{X}_{-1,(1)} = \boldsymbol{x}_1 \mathrm{vec}(\boldsymbol{A})^t + \exp\left\{\log(\boldsymbol{x}_1)\mathrm{vec}(\boldsymbol{B})^t\right\} * \boldsymbol{Z}_{|1,(1)}, \tag{E.2}$$

with $*$ the Hadamard product and $\boldsymbol{v}^t$ stands for the transpose of $\boldsymbol{v}$ considered as a single-column matrix. In terms of modes, we can rewrite (E.2) as

$$[R \times (d-1)S] = [R \times 1][(d-1)S \times 1]^t + [R \times 1][(d-1)S \times 1]^t * [R \times (d-1)S].$$

With the same idea we can compute the residual tensor $\mathcal{Z}_{|1}$ when estimating $\theta(x,m)$ through Monte Carlo integration. Let $\mathbf{1}_{(1)}$ denote the matricisation of a tensor of ones with modes $R \times (d-1) \times S$ and $\widetilde{x}$ the value at which $\theta(\widetilde{x},m)$ has to be estimated. We compute the residuals as follows:

$$\boldsymbol{Z}_{|1,(1)} = \left\{\mathbf{1}_{(1)}\widetilde{x} - \boldsymbol{x}_1 \mathrm{vec}(\boldsymbol{A})^t\right\} \not/ \exp\left\{\log(\boldsymbol{x}_1)\mathrm{vec}(\boldsymbol{B})^t\right\},$$

where $\not/$ stands for componentwise division.

F Confidence intervals for cluster maximum quantiles

For a given quantile level α, the corresponding quantile x satisfies

$$1 - r(x)\left(1 + \xi_{\text{all}}\frac{x-u}{\sigma_{\text{all}}}\right)^{-1/\xi_{\text{all}}} = \alpha, \quad \xi_{\text{all}} \neq 0, \tag{F.1}$$

$$1 - r(x)\exp\left(-\frac{x-u}{\sigma_{\text{all}}}\right) = \alpha, \quad \text{otherwise}, \tag{F.2}$$

where $r(x) := \theta(x,m)/\theta(u,m)$. The quantile function can then be computed based on $r = r(x)$:

$$q(r,\sigma_{\text{all}},\xi_{\text{all}}) := \begin{cases} \dfrac{\sigma_{\text{all}}}{\xi_{\text{all}}}\left\{\left(\dfrac{1-\alpha}{r}\right)^{-\xi_{\text{all}}} - 1\right\} + u, & \xi_{\text{all}} \neq 0, \\ -\sigma_{\text{all}}\log\left(\dfrac{1-\alpha}{r}\right) + u, & \text{otherwise.} \end{cases}$$

Assuming independence of $\widehat{r}$ with $(\widehat{\sigma}_{\text{all}}, \widehat{\xi}_{\text{all}})$ and asymptotic normality of the quantile estimator, we can get confidence intervals for the quantile estimator: applying a delta method, we obtain the quantile estimator variance

$$2\frac{\partial q}{\partial \sigma}\frac{\partial q}{\partial \xi}\text{cov}(\widehat{\sigma}_{\text{all}}, \widehat{\xi}_{\text{all}}) + \left(\frac{\partial q}{\partial \sigma}\right)^2 \text{var}(\widehat{\sigma}_{\text{all}}) + \left(\frac{\partial q}{\partial \xi}\right)^2 \text{var}(\widehat{\xi}_{\text{all}}) + \left(\frac{\partial q}{\partial r}\right)^2 \text{var}(\widehat{r}), \tag{F.3}$$

where $\widehat{r}$ is considered as being locally independent on x, as it varies slowly in x compared to the generalised Pareto distribution. This approximation allows us to compute the variance of $\widehat{r}$ easily. The derivatives in equation (F.3) are given by

$$\frac{\partial q}{\partial r}\left(\widehat{r},\widehat{\sigma}_{\text{all}},\widehat{\xi}_{\text{all}}\right) = \widehat{\sigma}_{\text{all}}\left(\frac{1-\alpha}{\widehat{r}}\right)^{-\widehat{\xi}_{\text{all}}-1}\frac{1-\alpha}{\widehat{r}^2},$$

$$\frac{\partial q}{\partial \sigma}\left(\widehat{r},\widehat{\sigma}_{\text{all}},\widehat{\xi}_{\text{all}}\right) = \begin{cases} \dfrac{1}{\widehat{\xi}_{\text{all}}}\left\{\left(\dfrac{1-\alpha}{\widehat{r}}\right)^{-\widehat{\xi}_{\text{all}}} - 1\right\}, & \xi_{\text{all}} \neq 0, \\ -\log\left(\dfrac{1-\alpha}{\widehat{r}}\right), & \text{otherwise,} \end{cases}$$

$$\frac{\partial q}{\partial \xi}\left(\widehat{r},\widehat{\sigma}_{\text{all}},\widehat{\xi}_{\text{all}}\right) = -\frac{\widehat{\sigma}_{\text{all}}}{\widehat{\xi}_{\text{all}}^2}\left\{\left(\frac{1-\alpha}{\widehat{r}}\right)^{-\widehat{\xi}_{\text{all}}} - 1\right\} - \frac{\widehat{\sigma}_{\text{all}}}{\widehat{\xi}_{\text{all}}}\left(\frac{1-\alpha}{\widehat{r}}\right)^{-\widehat{\xi}_{\text{all}}}\log\left(\frac{1-\alpha}{\widehat{r}}\right).$$

Computationally speaking, the ratios $\widehat{r}$ are only known on a predefined mesh of x's, and we find the solution to (F.1) by linearly interpolating $\widehat{r}$ between the mesh nodes. We use the same kind of linear approximation to get $\text{var}(\widehat{r})$, as the ratio variances vary smoothly throughout the mesh.

In equation (F.3), the variances and covariance for $\widehat{\sigma}_{\text{all}}$ and $\widehat{\xi}_{\text{all}}$ can be computed from the inverse Hessian matrix of the log-likelihood. However, to account for dependence between the exceedances, we apply an inflation method described by Smith (1990) and used by Fawcett and Walshaw (2007) in the same context as here; the

inverse Hessian H^{-1} is replaced by the product $H^{-1}VH^{-1}$. The matrix V is the covariance of independent score contributions from separate clusters. We write the log-likelihood function for a generalised Pareto distribution for n exceedances as

$$\ell(\sigma_{\text{all}},\xi_{\text{all}};\, x_1,\ldots,x_n) := \begin{cases} -n\log(\sigma_{\text{all}}) - \dfrac{1+\xi_{\text{all}}}{\xi_{\text{all}}}\sum_{i=1}^{n}\log\left(1+\xi_{\text{all}}\dfrac{x_i-u}{\sigma_{\text{all}}}\right), & \xi_{\text{all}} \neq 0, \\ -n\log(\sigma_{\text{all}}) - \sum_{i=1}^{n}\dfrac{x_i-u}{\sigma_{\text{all}}}, & \text{otherwise,} \end{cases}$$

and this can be expressed in terms of independent contributions:

$$\ell(\sigma_{\text{all}},\xi_{\text{all}};\, x_1,\ldots,x_n) = \sum_{c=1}^{C_m} \ell\left(\sigma_{\text{all}},\xi_{\text{all}};\, x_{c_1},\ldots,x_{c_{p(c)}}\right) =: \sum_{c=1}^{C_m} \ell_c,$$

with C_m the number of clusters corresponding to the run length m and observations $(x_{c_1},\ldots,x_{c_p})$ belong to cluster c. We denote by $p(c)$ the number of observations per cluster, which varies with c. Each estimated contribution $\widehat{\ell}_c$ is of the form:

$$\widehat{\ell}_c := \sum_{i=1}^{p(c)} \ell\left(\widehat{\sigma}_{\text{all}},\widehat{\xi}_{\text{all}};x_{c_1},\ldots,x_{c_{p(c)}}\right), \quad c = 1,\ldots,C_m,$$

thus giving an estimate for the correction matrix V:

$$\widehat{V} := \sum_{c=1}^{C_m} \left(\nabla\widehat{\ell}_c\right)\left(\nabla\widehat{\ell}_c\right)^t,$$

with score contributions having elements

$$\frac{\partial\widehat{\ell}_c}{\partial\sigma} = \sum_{i=1}^{p(c)} \frac{\dfrac{x_{c_i}-u}{\widehat{\sigma}_{\text{all}}} - 1}{\widehat{\sigma}_{\text{all}}\left(1+\widehat{\xi}_{\text{all}}\dfrac{x_{c_i}-u}{\widehat{\sigma}_{\text{all}}}\right)}, \quad c = 1,\ldots,C_m,$$

$$\frac{\partial\widehat{\ell}_c}{\partial\xi} = \sum_{i=1}^{p(c)} \frac{\left(1+\widehat{\xi}_{\text{all}}\dfrac{x_{c_i}-u}{\widehat{\sigma}_{\text{all}}}\right)\log\left(1+\widehat{\xi}_{\text{all}}\dfrac{x_{c_i}-u}{\widehat{\sigma}_{\text{all}}}\right) - \widehat{\xi}_{\text{all}}\left(1+\widehat{\xi}_{\text{all}}\right)\dfrac{x_{c_i}-u}{\widehat{\sigma}_{\text{all}}}}{\widehat{\xi}_{\text{all}}^2\left(1+\widehat{\xi}_{\text{all}}\dfrac{x_{c_i}-u}{\widehat{\sigma}_{\text{all}}}\right)}, \quad c = 1,\ldots,C_m.$$

G Capturing dependence of residuals

A first attempt to bring the one-step method into the multivariate framework was to make an inference on each conditional margin separately, thus implicitly assuming conditional independence of residuals. In this appendix we give a flavour of how the mixture of multivariate normal densities, even with independent margins, captures the dependence of the residuals.

We generated a 3-dimensional dataset from a Gumbel copula with dependence parameter $\zeta = 0.5$, in which case the Heffernan–Tawn parameters are $\boldsymbol{\alpha}_{|1} = 1$ and $\boldsymbol{\beta}_{|1} = 0$. The "true" residuals where computed using the data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_u}$, which have their first element above u, and the true values for the model parameters, i.e.,

$$Z_{i,j|1} = X_{i,j} - X_{i,1}, \quad i = 1, \ldots, n_u.$$

For the fitted models, the residuals are simply generated from the fits. They are sampled from the marginal mixtures of the model assuming them to be conditionally independent, and from the multivariate mixture for the model presented in the body of the text.

Figure 31 presents a comparison between the residuals derived from theoretical values, residuals generated under the conditional independence assumption, and under the conditional dependence assumption. The corresponding correlations are listed in Table 4, giving evidence in favour of assuming conditional dependence of the residuals.

	Pearson's r	Spearman's ρ	Kendall's τ
Theoretical	.7	.67	.49
Marginal fits	.12	.073	.049
Joint fit	.69	.66	.47

Table 4 – Correlation computed on the residuals for an example dataset of residuals derived from the theoretical values of the Heffernan–Tawn parameters (Theoretical), assumed conditionally independent (Marginal fits) and not assumed conditionally independent (Joint fit).

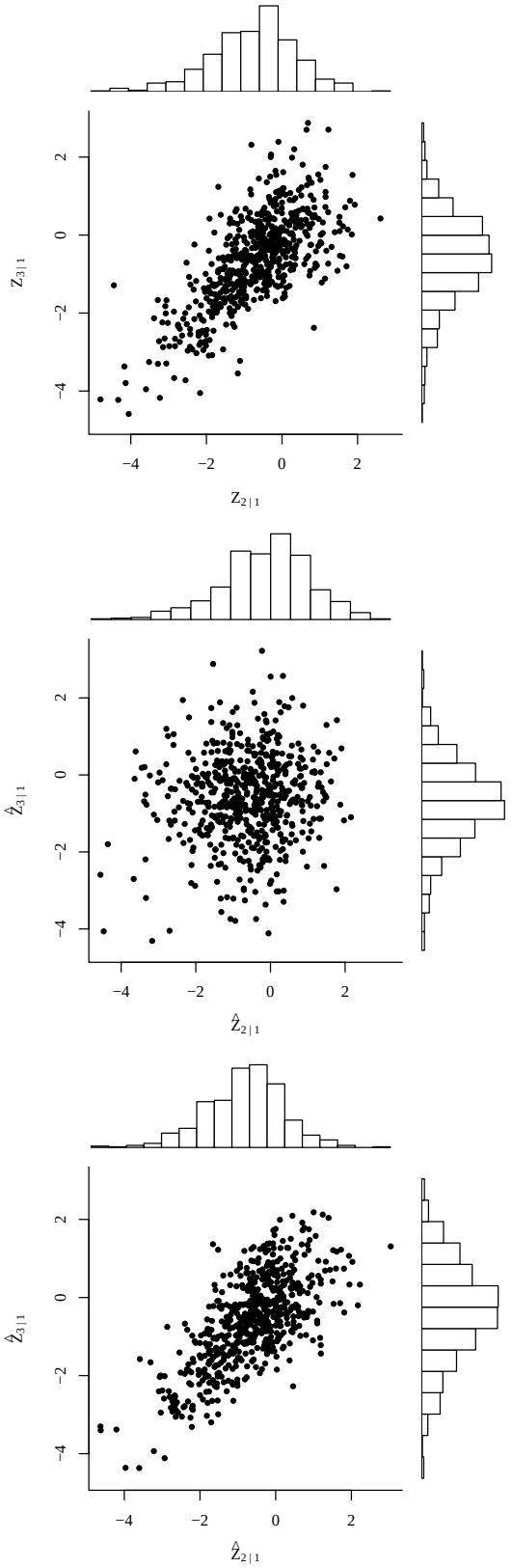

Figure 31 – Top panel: residuals derived from the true values for $\alpha_{j|1}$ and $\beta_{j|1}$, $j = 2,3$; middle panel: residuals generated from the model fitted separately on each margin; bottom panel: residuals generated from the model with multivariate mixture components.

References

Balkema, A. A. and Resnick, S. I. (1977) Max-infinite divisibility. *Journal of Applied Probability* **14**, 309–319.

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004) *Statistics of Extremes: Theory and Applications*. Chichester: Wiley.

Blackwell, D. and MacQueen, J. B. (1973) Ferguson distributions via Pólya urn schemes. *The Annals of Statistics* **1**, 353–355.

Brent, R. P. (1973) *Algorithms for minimization without derivatives*. New Jersey: Prentice Hall.

Breymann, W., Dias, A. and Embrechts, P. (2003) Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance* **3**, 1–16.

de Carvalho, M. and Ramos, A. (2012) Bivariate extreme statistics, II. *Revstat* **10**, 83–107.

Coles, S. G., Heffernan, J. E. and Tawn, J. A. (1999) Dependence measures for extreme value analyses. *Extremes* **2**, 339–365.

Coles, S. G. and Tawn, J. A. (1991) Modelling extreme multivariate events. *Journal of the Royal Statistical Society Series B* **53**, 377–392.

Coles, S. G. and Tawn, J. A. (1994) Statistical methods for multivariate extremes: An application to structural design (with discussion). *Journal of the Royal Statistical Society Series C* **43**, 1–48.

Connor, R. J. and Mosimann, J. E. (1969) Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association* **64**, 194–206.

Davison, A. C. and Smith, R. L. (1990) Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society Series B* **52**, 393–442.

Devroye, L. (1986) *Non-uniform random variate generation*. New York: Springer.

Eastoe, E. F. and Tawn, J. A. (2012) Modelling the distribution of the cluster maxima of exceedances of subasymptotic thresholds. *Biometrika* **99**, 43–55.

Escobar, M. D. (1994) Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89**, 268–277.

Fawcett, L. and Walshaw, D. (2007) Improved estimation for temporally clustered extremes. *Environmetrics* **18**, 173–188.

Fearnhead, P. (2004) Particle filters for mixture models with an unknown number of components. *Statistics and Computing* **14**, 11–21.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

Fermanian, J.-D., Radulović, D. and Wegkamp, M. (2004) Weak convergence of empirical copula processes. *Bernoulli* **10**, 847–860.

Gumbel, E. J. and Goldstein, N. (1964) Analysis of empirical bivariate extreme distributions. *Journal of the American Statistical Association* **59**, 794–816.

Gustard, A., Bullock, A. and Dixon, J. M. (1992) *Low Flow Estimation in the United Kingdom*. Institute of Hydrology.

de Haan, L. and de Ronde, J. (1998) Sea and wind: Multivariate extremes at work. *Extremes* **1**, 7–45.

Heffernan, J. E. (2000) A directory of coefficients of tail dependence. *Extremes* **3**, 279–290.

Heffernan, J. E. and Tawn, J. A. (2004) A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society Series B* **66**, 497–546.

Hill, B. M. (1975) A simple general approach to inference about the tail of a distribution. *The Annals of Statistics* **3**, 1163–1174.

Institute of Hydrology (1980a) Low Flow Studies. Report No 1: Research Report.

Institute of Hydrology (1980b) Low Flow Studies. Report No 3: Catchment Characteristic Estimation Manual.

Ishwaran, H. and James, L. F. (2001) Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–172.

Ishwaran, H. and James, L. F. (2002) Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics* **11**, 508–532.

Ishwaran, H. and Zarepour, M. (2000) Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87**, 371–390.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005) Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science* **20**, 50–67.

Jonathan, P., Ewans, K. and Randell, D. (2013) Joint modelling of extreme ocean environments incorporating covariate effects. *Coastal Engineering* **79**, 22–31.

Keef, C., Papastathopoulos, I. and Tawn, J. A. (2013) Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model. *Journal of Multivariate Analysis* **115**, 396–404.

Keef, C., Tawn, J. and Svensson, C. (2009) Spatial risk assesment for extreme river flows. *Journal of the Royal Statistical Society Series C* **58**, 601–618.

Kiers, H. A. L. (2000) Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics* **14**, 105–122.

Kolda, T. G. and Bader, B. W. (2009) Tensor decompositions and applications. *Society for Industrial and Applied Mathematics Review* **51**, 455–500.

Leadbetter, M. R. (1983) Extremes and local dependence in stationary sequences. *Probability Theory and Related Fields* **65**, 291–306.

Ledford, A. W. and Tawn, J. A. (1996) Statistics for near independence in multivariate extreme values. *Biometrika* **83**, 169–187.

Ledford, A. W. and Tawn, J. A. (2003) Diagnostics for dependence within time series extremes. *Journal of the Royal Statistical Society Series B* **65**, 521–543.

MacEachern, S. N. and Müller, P. (1998) Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–238.

Marsh, T. J. and Hannaford, J. (2008) UK Hydrometric Register. *Hydrological Data UK Series. Centre for Ecology and Hydrology* .

Neal, R. M. (2000) Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

Nelder, J. A. and Mead, R. (1965) A simplex method for function minimization. *The Computer Journal* **7**, 308–313.

Papaspiliopoulos, O. and Roberts, G. O. (2008) Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**, 169–186.

Pickands, J. (1981) Multivariate extreme value distributions. In *Bulletin of the International Statistical Institute, Proceedings of the 43rd Session*, volume 2, pp. 859–878.

Porteus, I., Ihler, A., Smyth, P. and Welling, M. (2006) Gibbs sampling for (coupled) infinite mixture models in the stick breaking representation. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*.

Postman, M., Huchra, J. P. and Geller, M. J. (1986) Probes of large-scale structure in the Corona Borealis region. *The Astronomical Journal* **92**, 1238–1247.

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Redner, H. F. and Walker, R. A. (1984) Mixture densities, maximum likelihood and the EM algorithm. *Society for Industrial and Applied Mathematics Review* **26**, 195–239.

Resnick, S. I. (1987) *Extreme Values, Regular Variation, and Point Processes*. New York: Springer.

Robert, C. P. (1995) Simulation of truncated normal variables. *Statistics and Computing* **5**, 121–125.

Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association* **85**, 617–624.

Schlather, M. and Tawn, J. A. (2003) A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika* **90**, 139–156.

Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Smith, R. L. (1990) Regional estimation from spatially dependent data. Unpublished.

Smith, R. L. and Weissman, I. (1994) Estimating the extremal index. *Journal of the Royal Satistical Society Series B* **56**, 515–528.

Stephens, M. (2000) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society Series B* **62**, 795–809.

Tawn, J. A. (1988) Bivariate extreme value theory: Models and estimation. *Biometrika* **75**, 397–415.

Tucker, L. R. (1966) Some mathematical notes on three-mode factor analysis. *Psychometrika* **31**, 279–311.

Venables, W. N. and Ripley, B. D. (2002) *Modern applied statistics with S*. New York: Springer.